T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF BIG DATA ANALYTICS AND MANAGEMENT

USE OF MACHINE LEARNING REGRESSION ALGORITHMS FOR SALARY PREDICTION IN TURKISH JOB MARKET

MASTER'S THESIS

KARIM ALDEIK

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF BIG DATA ANALYTICS AND MANAGEMENT

USE OF MACHINE LEARNING REGRESSION ALGORITHMS FOR SALARY PREDICTION IN TURKISH JOB MARKET

MASTER'S THESIS

KARIM ALDEIK

THESIS ADVISOR
Asst. Prof. Dr. SELAMI BAGRIYANIK

ISTANBUL 2024

T.C.

BAHÇEŞEHİR UNIVERSITY

GRADUATE SCHOOL

...../...../.....

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics and Management
Student's Name and Surname:	Karim Aldeik
Name of The Thesis:	Use of Machine Learning Regression Algorithms for Salary Prediction in Turkish Job Market
Thesis Defense Date	16/01/2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Dr. Yücel Batu Salman

Institute Director

This thesis was read by us, quality and content as a master's thesis has been seen and accepted as sufficient.

	Title, Name	**Institution**	**Signature**
Thesis Advisor:	Asst. Prof. Dr. Selami Bagriyanik	Topkapi University - Faculty of Engineering	
2nd Member	Asst. Prof'. Dr. Tamer Uçar	BAU Faculty of Engineering and Natural Sciences	
3rd Member (Outside Institution)	Asst. Prof. Dr Suat Uğurlu	Topkapi University - Faculty of Engineering	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as requiredby these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Karim Aldeik

Signature

ABSTRACT

USE OF MACHINE LEARNING REGRESSION ALGORITHMS FOR SALARY PREDICTION IN TURKISH JOB MARKET

Karim Aldeik
Big Data Analytics and Management Master's Program
Thesis Advisor: Asst. Prof. Dr. SELAMI BAGRIYANIK

January 2024, 47 pages

In today's dynamic job market, accurately predicting salaries is a critical task for both employees and employers. Salary prediction can help employees negotiate fair compensation for their skills and experience, and help employers attract and retain qualified talent. However, salary prediction is not a straightforward problem, as it depends on various factors such as job titles, skills, education, location, industry, and company size. This study delves into the realm of salary prediction by leveraging essential factors such as job titles and skills within a dataset. The dataset collected from a popular job search website. The dataset contains information about the job title, skills needed, industry, and salary of each job posting. Job titles provide insight into an individual's role and responsibilities, while skills reflect their expertise and proficiency. Harnessing the power of these features, this research employs diverse machine learning models to predict the salary range of a given job posting. The results indicate promising performance across the models, with the achievement of a noteworthy 6.5 Mean Absolute Percentage Error (MAPE) in the skill score method. The study provides valuable insights into the effectiveness of machine learning models in salary prediction, emphasizing the significance of incorporating job titles and skills for enhanced accuracy. These findings contribute to the ongoing discourse on optimizing salary prediction methodologies in the evolving landscape of the job market.

Keywords: Salary Prediction ,Machine Learning ,Job Titles, Skills ,Human Resources

ÖZ

USE OF MACHINE LEARNING REGRESSION ALGORITHMS FOR SALARY PREDICTION IN TURKISH JOB MARKET

Karim Aldeik

Big Data Analytics and Management Master's Program
Tez Danışmanı: Asst. Prof. Dr. SELAMI BAGRIYANIK

Ocak 2024, 47 sayfa

Günümüz dinamik iş piyasasında, maaşların doğru bir şekilde tahmin edilmesi, hem çalışanlar hem de işverenler için kritik bir konudur. Maaş tahmini, çalışanların yetenek ve deneyimleri için adil bir müzakere yapmalarına yardımcı olabilirken, işverenlere nitelikli yetenekleri çekme ve elde tutma konusunda destek sağlayabilir. Ancak, maaş tahmini çözümü basit bir problem değildir, çünkü iş unvanları, beceriler, eğitim, konum, endüstri ve şirket büyüklüğü gibi çeşitli faktörlere bağlıdır. Bu çalışma, bir veri kümesi içinden iş unvanları ve beceriler gibi temel faktörleri kullanarak maaş tahminine yönelik bir alanı incelemektedir. Popüler bir iş arama web sitesinden toplanan veri kümesi, her iş ilanının iş unvanı, gereken beceriler, endüstri ve maaş bilgilerini içermektedir. İş unvanları, bir bireyin rol ve sorumluluklarına dair içgörü sağlarken, beceriler ise uzmanlıklarını ve yeteneklerini yansıtmaktadır. Bu özelliklerin gücünden faydalanarak, bu araştırma çeşitli makine öğrenimi modellerini kullanarak belirli bir iş ilanının maaş aralığını tahmin etmektedir. Sonuçlar, modeller arasında umut verici bir performans göstermekte olup, beceri puanı yönteminde dikkat çekici 6.5 Ortalama Mutlak Yüzde Hata (MAPE) başarısını göstermektedir. Bu sonuçlar, iş piyasasının gelişen durumunda maaş tahmini metodolojilerinin optimize edilmesine dair devam eden tartışmalara değerli bir katkı sağlamaktadır. Çalışma, iş başlıkları ve becerilerin entegrasyonunun daha yüksek doğrulukla öngörü için ne kadar önemli olduğunu vurgulayarak, makine öğrenimi modellerinin etkinliği konusunda değerli içgörüler sunmaktadır.

Anahtar Kelimeler: Maaş Tahmini, Makine Öğrenimi, Iş Unvanları, Beceriler, Insan Kaynakları.

To The Resilient and Courageous People of Palestine, who strive for freedom and justice, May the pursuit of knowledge and the power of science contribute to the path towards liberation, May education be a beacon of hope and pave the way for a future of peace and prosperity.

ACKNOWLEDGMENTS

I extend my sincere gratitude to my supervisor Asst. Prof. Dr. Selami Bagriyanik my advisor, whose guidance and expertise have been invaluable throughout the course of this research. Your encouragement, insights, and your support have played a crucial role in shaping the direction of this thesis.

I would also like to express my deepest appreciation to my parents. Your love, encouragement, and sacrifices have been the foundation of my academic journey. Your unwavering belief in my abilities has been a constant source of motivation.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ERM	Empirical Risk Minimization
IKL	Infinite Kernel Learning
MEB	Minimum Enclosing Ball
MKL	Multiple Kernel Learning
QCQP	Quadratically Constraint Quadratic Programming
QP	Quadratic Programming
RBMK	Rule Based Multiple Kernel
SMO	Sequential Minimal Optimization
SLT	Statistical Learning Theory
SRM	Structural Risk Minimization
SV	Support Vector
SVboosting	Support Vector Boosting
SVM	Support Vector Machine
VC-dim	Vapnik Chervonenkis dimension
VC	Vapnik Chervonenkis
KNN	K- Nearest Neighbors

Chapter 1

Introduction

1.1 Statement of the Problem

In today's competitive job market, accurately determining salaries is paramount for both job seekers and employers. Predicting salaries assists organizations in promoting fairness and equity, making well-informed compensation decisions, and luring and keeping top talent (Gurpreet et al., 2023). However, this task is complex due to the multitude of factors influencing salaries, including experience, education, location, and skills. Understanding the correlation between specific attributes, such as job titles and skills, and the corresponding salaries is essential. This study aims to address the challenge of salary prediction by focusing on the utilization of job titles and skills as key features. By employing a range of machine learning models, the research seeks to unravel the intricate relationship between these features and salary levels. The primary problem lies in identifying the most effective machine learning model for accurate salary predictions while ensuring a comparative analysis to discern the strengths and weaknesses of each approach. Additionally, this research endeavors to contribute to the advancement of salary prediction techniques, thereby aiding job seekers in negotiating competitive compensation and assisting employers in making informed hiring and remuneration decisions.

1.2 Purpose of the Study

The primary purpose of this study is to develop a robust and reliable salary prediction model using essential features such as job titles and skills derived from a comprehensive dataset. Artificial intelligence and machine learning have made it possible for complex algorithms to evaluate vast amounts of data and predict salaries with high accuracy (Eichinger & Mayer, 2022). Unfortunately, one of the challenges that face salary prediction is the scarcity of large real datasets for learning.

(Mallapaty, 2018), Another challenge is the necessity to consider multiple factors when predicting salaries making the task complex. the large number of variables can affect the accuracy predictions highlighting the importance of reducing the number of values for these variables (Goerke & Neugart, 2020). Outlier handling is also crucial for achieving high-quality salary predictions. this research endeavors to construct predictive models that can accurately forecast salaries based on job titles and skill sets by harnessing the potential of machine learning,

Additionally, this study aims to conduct a comparative analysis of multiple machine learning models to identify the most effective approach for salary prediction. The overarching goal is to empower both job seekers and employers with an advanced tool that enhances transparency in salary negotiations and facilitates informed decision-making. By achieving this purpose, the study contributes to bridging the information gap in salary expectations and, in turn, fosters a more equitable and efficient job market.

1.3 Research Question and Hypothesis

This study delves into the following fundamental research question: Can a predictive model, utilizing job title and skill-related features, accurately predict salaries in various job markets? To address this question, we hypothesize that machine learning models, when trained on a comprehensive dataset encompassing job titles and skills, can yield accurate salary predictions. Furthermore, we posit that different machine learning algorithms may exhibit varying levels of predictive accuracy and efficiency in salary estimation. By formulating and testing this hypothesis, we aim to unravel the potential of machine learning in salary prediction and identify the optimal models for this task. Ultimately, this research endeavors to shed light on the effectiveness of leveraging data-driven approaches to enhance salary prediction and contribute to more transparent and equitable employment practices.

1.4 Significance of the Study

Salary prediction studies in Turkey are significant for both job seekers and employers. This significance is supported by the multitude of studies that have focused on this subject in the past years, as evidenced by (Özer et al., 2022) And (Şentürk, 2015). and wage inequality problem has been rising in the recent years (Tansel et al., 2019) therefore the need for a salary reference point for negotiation is very helpful.

This study holds significant implications for both academia and industry. In the academic realm, it contributes to the growing field of machine learning and data analytics by investigating an essential socio-economic concern: salary prediction. By utilizing job titles and skills as predictive features, this research advances the understanding of how these fundamental variables influence salary outcomes. Additionally, the comparative analysis of multiple machine learning models enriches the knowledge base regarding the applicability and effectiveness of these algorithms in the context of salary prediction. From an industry perspective, the insights garnered from this study can guide human resource management and policy decisions. Accurate salary prediction is critical for fair compensation and can aid organizations in attracting and retaining talent, optimizing budget allocation, and fostering a transparent and equitable work environment. Ultimately, this study endeavors to bridge the gap between academic research and practical industry applications, promoting informed decision-making in the realm of workforce compensation.

Chapter 2

Literature Review

The setting of fair and competitive compensation plays a crucial role in luring and keeping talented individuals in today's dynamic and competitive employment market. Employers and HR experts are increasingly using cutting-edge data-driven strategies, especially Artificial Intelligence (AI) and neural network models, to forecast and determine the proper pay scales for employees. This creative application of AI guarantees that compensation is in line with market norms, individual credentials, and job functions while also streamlining salary management. In the fields of human resources (HR) and organizational management, the convergence of AI and wage prediction has become a crucial topic for research and application.

Accurate salary prediction based on a variety of employee-related data has the potential to completely alter the HR environment. It enables businesses to make educated decisions about worker remuneration, promotes fair pay practices, and ultimately raises employee retention and satisfaction. Additionally, the use of AI and neural network models in pay prediction offers chances to scale these procedures, lessen human bias, and adjust to the always-changing nature of the labor market.

This literature review explores the complex area of compensation prediction using employee data. It attempts to offer a thorough review of the state of the field's research, methodologies, difficulties, and new trends. We aim to acquire insights into the methodology used, the parameters considered, the accuracy attained, and the practical consequences of AI-driven salary prediction by examining the corpus of existing research.

The paper (Saito et al., 2019) introduces TurkScanner, a machine learning methodology designed to predict both the duration and hourly wage associated with crowd-work assignments. The researchers developed a technique for gathering data for practical training and examined the difficulty of precisely documenting working hours. TurkScanner aims to address underpayment and unfair treatment of workers in crowd

markets by predicting hourly wages for microtasks. Future research should focus on empowering workers and requesters and evaluating how tools affect working processes.

This paper aims to address underpayment and unfair treatment of workers in crowd markets by developing a machine-learning method that estimates the amount of time and money needed each hour to do microtasks. The objective is to accurately record working times, both automatically and by asking workers. Future directions for this research include leveraging TurkScanner to assist requesters' and workers' empowerment tools and tracking how these tools affect people's work practices.

The data used in this paper consists of 9,155 HIT ("Human Intelligence Task," which refers to a small, simple task that a human can complete quickly and easily, such as labeling an image or transcribing a short audio clip.) Submitted data gathered by 84 individuals. 998 distinct requesters posted 1,641 unique HIT groups, which included the recorded HITs. The data was collected for 10 days during late October 2018.

This paper used a Google Chrome extension to collect data on AMT ("Amazon Mechanical Turk," which is a popular platform for crowdsourcing microtasks.) workers' working hours. The extension automatically recorded working time, provided a button for manual recording and asked workers to select the best time for labeling. A pre-survey was prepared to recruit workers, asking for profile information like gender, country, age, income, experience, and hours per week. Low-quality data was filtered to ensure the dataset's quality.

The paper discusses the challenges faced in accurately recording working times in a microtasking system. The researchers developed a Google Chrome extension to collect data but found that some workers did not use it or did not accurately report their working time. They also had to filter out low-quality data, including spam workers using automated scripts, abnormally short or long working times, and a large portion of the top three most-submitted HIT groups. Additionally, some types of HITs, such as survey HITs with external URLs and dynamically rendered microtasks, were difficult to predict accurately due to their lack of static HTML content.

TurkScanner, a tool for predicting hourly wage, achieved a 75% working-time error for 69.6% of microtasks and 100% error for 84.3%. It performed well for HIT records with actual hourly wages below $15 USD, but many higher wages were not predicted. The study had limitations, including difficulty recording working times and predicting HITs with external URLs and dynamically rendered content.

TurkScanner is a valuable tool for workers and requesters in crowded markets, allowing them to estimate hourly wages and accept tasks, while requesters can set fair wages and attract more workers. Future research should focus on empowering workers and requesters and observing the impact of these tools on working practices.

The paper discusses TurkScanner, a tool that aids workers in deciding which microtasks to start and assist requesters in price setting. It introduces an innovative technique for figuring out and predicting the time arbitrary user interface tasks will take, which could be used in various settings like determining wages, organizing time, or automated usability testing. The authors analyze key features for predicting working time, ensuring reliability. They recommend using TurkScanner in future studies to study how tools affect productivity and to empower workers and requesters. The paper's authors highlight several limitations of their study, including the difficulty in accurately recording working times, predicting HITs with external URLs and dynamically rendered content, and the potential for the dataset to not represent all microtasks on crowd markets. They also acknowledge that TurkScanner is not a perfect solution and that there is still room for improvement in terms of accuracy and usability. The authors acknowledge that their study's limitations highlight the need for further research and improvement.

This paper (Jafari, Rouhanizadeh, Kermanshachi, & Murrieum, 2020) presents a framework for using predictive analytics to evaluate wage inequality in engineering organizations. The framework includes a salary prediction model, a minority index, and a decision model. The study found that even though women are underrepresented in USDOT (United States Department of Transportation. It is a federal government agency responsible for overseeing and regulating transportation in the United States, including highways, aviation, railroads, and maritime transportation.), There is no notable disparity

in earnings between males and females. The paper also discusses the potential limitations of using predictive analytics to evaluate wage inequality.

This paper aims to use predictive analytics to assess pay disparities in engineering firms. This research seeks to establish a robust system for employers to determine equitable wage rates, free from prejudice based on gender, ethnicity, or disability. The project targets the development of three key tools: a salary prediction model, a minority index, and a decision-making framework. These tools aim to identify employees susceptible to wage disparity and assess the current state of diversity and wage fairness within an organization.

This study's research methodology involved several steps:

To perform an in-depth evaluation of the organization's current workforce, investigating both the representation of diverse groups and the presence of potential wage disparities, data collection and analysis were first carried out.

Second, a logistic regression model was used to analyze the effects of correlated variables on employee wages to investigate wage inequality among workers. A salary range prediction model's correlated predictor variables were also made clearer thanks to this analysis.

Third, the diversity status of USDOT employees was examined using the related rate index (RRI).

Finally, a framework that employs machine learning techniques to find workers at risk of wage inequality and establish an objective wage rate has been developed. A decision model, a minority index, and a salary prediction model are all included in the framework.

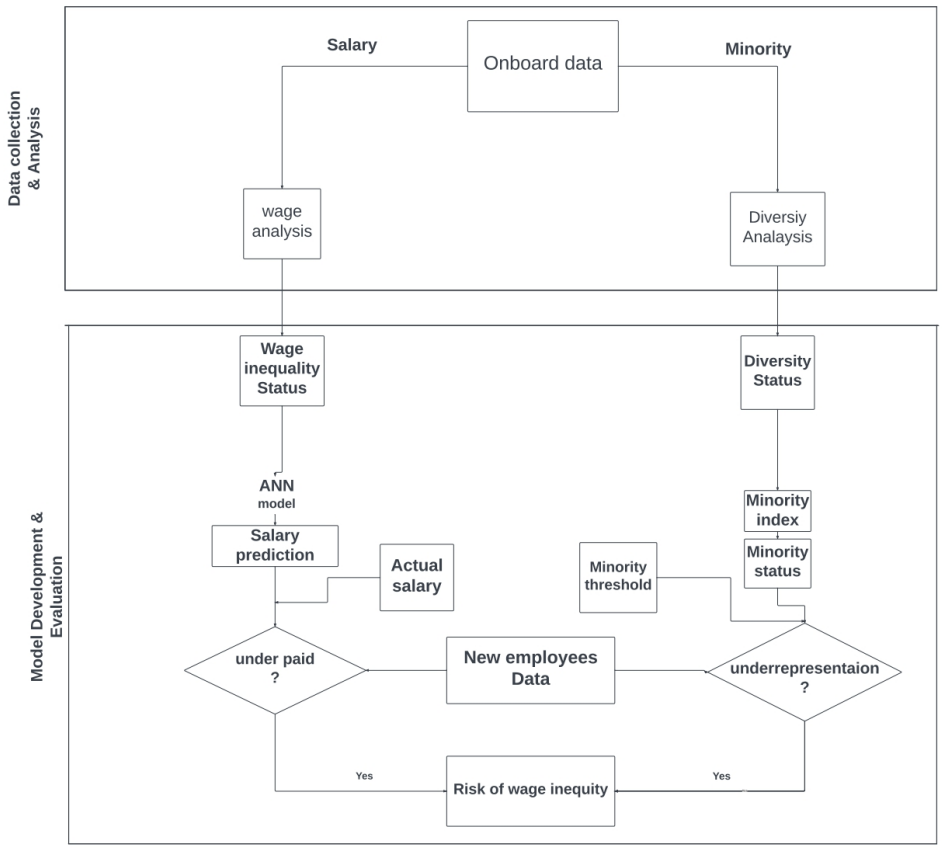

Figure 1. The research methodology Diagram

(Jafari, A, 2020)

The research leveraged workforce demographic and salary data spanning 2011 to 2017, provided by the USDOT, encompassing a total of 17,889 employees. The proposed framework required three types of data: minority status, salary range, and salary predictors. Salary predictors are variables evaluated for salary calculation, such as education level, job position, experience level, and years of service. The data collection method is not explicitly stated, but it is likely obtained through internal records or surveys conducted by USDOT.

Figure 2. Descriptive statistics of salary range variable.

(Jafari, A,2020)

The paper discusses the limitations and difficulties of predictive analytics in evaluating wage inequality. The main challenge is the limited availability and quality of data, which may not be generalizable to other organizations. The framework used is also susceptible to bias in the data or the algorithms used to analyze it. The framework is also limited to evaluating wage inequality based on demographic factors and does not consider other factors contributing to wage inequality, such as job performance or negotiation skills. These challenges highlight the need for more comprehensive and accurate data analysis in wage inequality evaluation.

The research proposes a system capable of recognizing employees who may face wage inequality and establishing a fair compensation rate, irrespective of gender, ethnicity, or disability status. The framework accurately predicted wage inequality in an engineered data set, with a high true positive rate and a low false positive rate. The study concludes that machine learning methods can be a useful tool for addressing wage inequality prediction and policy-making problems in engineering organizations.

The study suggests a framework for employers to create a more equitable workplace and reduce wage inequality's negative effects, such as job satisfaction and turnover. It suggests applying the framework to large data sets and evaluating wage

inequality at the organizational level. The study emphasizes the use of machine learning methods to address wage inequality issues.

The paper addresses wage inequality in the workplace, a significant concern for organizations and employees. It proposes a new framework using machine learning methods to identify employees at risk and create an unbiased wage rate. The study's transparency and reproducibility are enhanced by detailed descriptions of data and analysis methods. The paper also discusses the limitations and challenges of predictive analytics in evaluating wage inequality. The proposed framework's conclusions and implications are clear and concise, useful for employers and policymakers in addressing wage inequality in the workplace.

This paper (Dutta, Halder, & Dasgupta, 2018) presents a proposed model for salary prediction using machine learning techniques. The proposed Prediction Engine can predict the salary of a job based on key features such as Title, contract type and time, location, company, category, job description, source name. The paper discusses the preprocessing steps involved with the raw data set, data processing, and the proposed Prediction Engine. The results obtained are encouraging and with high precision. The paper concludes with future scope for the proposed model.

The research question of this paper is how to automate and create a proposed model for salary prediction using machine learning techniques. The aim of the paper is to develop a Prediction Engine that can predict the salary of a job based on key features. The objective of the paper is to improve the experience of people searching for jobs and help employers and job seekers to determine the current market value for various positions.

The dataset used in this paper is collected and uploaded to www.kaggle.com by ADZUNA (a company that collected and uploaded the dataset), from where the authors have downloaded the dataset. The dataset contains 244,768 data points and includes information such as job description, title, location, contract type, contract time, company, category, and source name. The authors have divided the dataset into two parts - train

dataset and test dataset, with a divide ratio of 1/3 in the test set and 2/3 in the train set. The data collection method is not explicitly mentioned in the paper.

The paper discusses the challenges of salary prediction, highlighting the need for careful consideration of factors like qualifications, work experience, project number, location, category, and company. It also highlights the need for managers to analyze large amounts of data before making decisions, making it impossible for humans to make reliable decisions. The authors suggest that they may have faced difficulties in data analysis and modeling, suggesting that a model is needed to intelligently handle this task.

The paper presents a Prediction Engine using machine learning techniques like decision tree and random forest classifiers to predict job salaries with high precision. The model considers job description, title, location, contract type, contract time, company, and category. It improves job search experiences and helps employers and job seekers understand market worth for different positions. The model can be used by companies and job portals to predict job salaries, provide better job recommendations, and help companies decide fair salaries and attract the right talent.

The strong points of the paper are:

1. This paper introduces a novel Prediction Engine employing machine learning algorithms like decision trees and random forests. Leveraging key job details such as description, title, location, contract specifics, company, and category, the engine delivers accurate salary predictions for various positions.
2. 2. The experiments carried out on the suggested model are shown in the paper, demonstrating the model's high degree of accuracy in predicting an employee's salary.
3. The paper discusses the implications of the proposed model for job seekers, employers, and job portals, which can help to improve the job search experience and attract the right talent.

Regarding the weak points of the paper, there are a few things to consider:

1. The paper does not provide a detailed explanation of the dataset used in the experiments, such as the size of the dataset, the sources of the data, and the criteria used for selecting the data.
2. The paper does not compare the proposed model with other existing models or methods for salary prediction, which makes it difficult to evaluate the effectiveness of the proposed model in comparison to other approaches.
3. The paper does not provide a detailed explanation of the interpretability of the proposed model, which is important for understanding how the model makes predictions and for building trust in the model among stakeholders.

Overall, while the proposed model shows promising results, the lack of information on the dataset, comparison with other models, limitations, and interpretability of the model are some of the weak points that need to be addressed in future research.

This paper (Quan & Raheem, 2022) is a literature review that explores the latest methodologies for creating a more accurate salary prediction model in the data science field. It identifies common human resource problems and the most in-demand skills in this industry. The paper also explores the influence of job benefits on salary predictions and classifies machine learning methodologies for benchmarking into three primary categories, each exhibiting specific strengths based on diverse situations and requirements. This categorization is approached through statistical methods, ensemble machine learning methods, and deep learning-based neural networks are discussed in detail.

This paper aims to develop a salary prediction model for data science professionals based on specialized skills and job benefits. It reviews existing methodologies and addresses human resource issues in the field. The objectives include identifying the most demanded skillset for early exploration, predicting accurate salaries involves projecting an experimental dataset that incorporates factors such as skills and job benefits. The process also entails the careful selection of appropriate methodologies for constructing a salary prediction model in the field of data science. The paper does not provide specific information on data collection methods or data sources. Nevertheless, it is stated that the experimental dataset was developed to encompass factors related to skills and job benefits,

aiming to enhance the precision of salary predictions. The paper also discusses the importance of accurate recruitment in improving an organization's market competitiveness and work productivity, which suggests that data on job titles and skills may have been collected from job postings or surveys of employers in the data science field.

The paper does not explicitly mention difficulties faced in the research process. However, While the paper introduces an initial framework for the data science salary prediction model, it acknowledges the need for further research to solidify the project's objectives, scenario, and data selection. Additionally, it expresses the potential for a deeper investigation into data science job benefits, contingent upon the availability of sufficient studies. These statements suggest that the research team may have faced challenges in identifying appropriate research objectives and selecting suitable data sources for the study.

The research outcomes encompass recognizing the most sought-after data science skill sets for feature selection. Additionally, the paper evaluates established methodologies for constructing a salary prediction model, considering specialized skills and specific job benefits. It further outlines appropriate benchmarking models utilizing various techniques. It highlights crucial data mining, visualization, and analysis processes for successfully developing the model and achieving business objectives.

Regarding suitable benchmarking models with different techniques, the paper categorizes the reviewed benchmarking categorizing machine learning methodologies into three primary groups, each possessing unique strengths tailored to diverse situations and requirements. These categories are:

1. Statistical methods: These methods excel at illustrating relationships between variables with extraordinary parameter tuning potential if linearity is present.

2. Ensemble machine learning methods: These methods, like Random Forest, combine multiple classifiers for more stable and accurate prediction.

3. Deep learning-based neural networks: These methods have a strong specialty in handling unlabeled data and framework modifications.

The paper conducts a comparative analysis between traditional statistical or parametric models and contemporary non-parametric models, particularly focusing on ensemble tree-based models. Evaluated parametric models encompass Quantile Regression, Linear Regression, and Logistic Regression, with Random Forest and Gradient Boosting serving as non-parametric models. In both classification and regression tasks, the experimental results indicate that the Gradient Boosting model outperformed the others, with Random Forest and Regression following in performance.

In the conclusion section of the paper, the key findings are summarized, and it is proposed that a salary prediction model, utilizing specialized skills and relevant job benefits, has the potential to effectively address challenges and fulfill the needs within the field of data science. The paper further highlights that this predictive tool could serve as both a motivator and guide for students, encouraging them to pursue studies and skills in specific areas. This, in turn, is seen as a step towards alleviating the identified issue of the shortage of data science jobs in the United States.

The paper underscores that precise recruitment stands as a pivotal element in enhancing an organization's market competitiveness and work productivity, considering employees as the company's most valuable asset. It posits that constructing a predictive model for salary determination, incorporating specialized skills and job benefits, holds the potential to address human resource challenges in the field of data science and enhance overall business objectives and achievements. Additionally, the paper emphasizes the significance of data mining, visualization, and analysis processes in ensuring the effectiveness of model building.

The paper has several strong points, including:

1. Comprehensive review: The paper conducts an extensive review of established methodologies for constructing a salary prediction model, focusing on specialized skills and specific job benefits within the data science domain. The assessment of benchmarking machine learning methodologies related to these issues results in their categorization into three primary groups, each exhibiting distinct strengths suited for various situations and requirements.

2. Clear presentation: The paper presents the information in a clear and concise manner, making it easy to understand for readers who are not experts in the field.

3. Practical implications: The paper highlights the practical implications of building a salary prediction model based on specialized skills and given job benefits, including resolving human resource problems in the data science field and improving business objectives achievements.

4. Research gaps: The paper identifies research gaps in the field, such as the need for more studies on data science-related job benefits and Additional efforts are required to define the project's goals, scenario, and experimental dataset to choose appropriate methodologies for constructing a data science salary prediction model.

5. Implications for education: The paper proposes that a salary prediction model, grounded in specialized skills and relevant job benefits, could serve as both a motivator and guide for students. This approach has the potential to encourage students to pursue studies and develop skills in specific fields, thereby contributing to the gradual alleviation of the identified issue of a shortage of data science jobs in the United States.

This paper (Görmez et al., 2022), published in Advances in Artificial Intelligence Research (AAIR) Vol. 2 (No. 1) (The (AAIR) Vol. 2 (No. 1) refers to the second volume and first issue of the Advances in Artificial Intelligence Research (AAIR) journal. It is a publication that includes research articles related to artificial intelligence and its applications. The journal is published online, and the specific issue was published on February 16, 2022.), presents a machine learning-based system design to predict salary increases. The study analyzes the relationship between personnel salary, motivation, and productivity, and concludes that salary increase rate estimation can be made with machine learning methods. The designed system can be integrated into ERP systems and uses a study-specific scale to calculate performance scores. The system is crafted to align with future corporate requirements, employing a microservice architecture and flexible structures enabling the machine learning model to autonomously update. The authors intend to implement the proposed system within a company within the information

technology sector for testing and evaluation. The paper also includes a literature review on the effects of salary on individuals.

Whether it is possible to forecast pay rise rates using machine learning techniques is the research issue addressed in this work. The study's objective is to develop an ERP system (ERP stands for Enterprise Resource Planning. It is a type of software that companies use to manage their business processes, such as accounting, procurement, and human resources) machine learning-based system that can forecast wage rise rates. The study's goals include examining the connection between employee pay, motivation, and output; creating a method for calculating performance scores; creating a machine learning-based system that can forecast salary increase rates; and testing the system in an IT company.

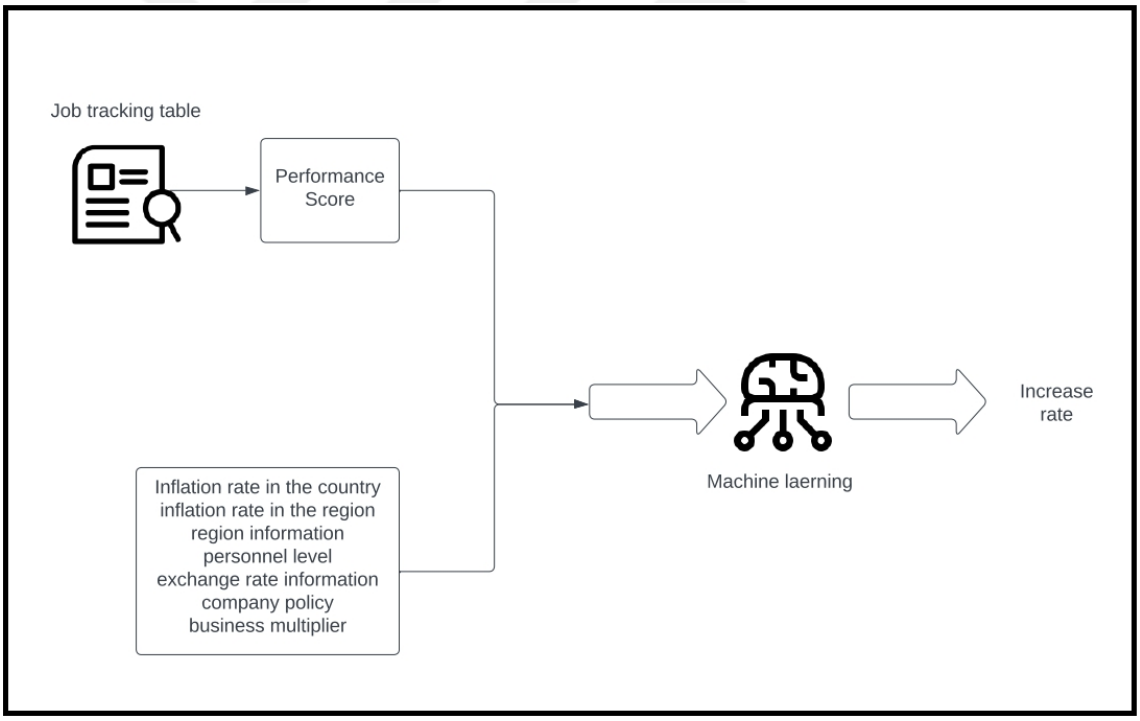

Figure 3. Proposed system architecture

((Görmez et al., 2022),

The paper does not provide detailed information about the data collection method, or the specific data used in the study. However, it is mentioned that data from consulting

firms such as Korn Ferry (Hay Group) that collected market analysis data because of surveys were used to create the features.

Any challenges encountered throughout the study are not specifically mentioned in the publication. The authors do, however, state that a micro-service architecture will be used in the system's development to modularize it. Different machine learning models can be created in this way for various businesses. If the new company desires to utilize the technology, a tailored model will be trained using its own historical data to ensure company-specific accuracy. This implies that one potential challenge in creating the system is the requirement to tailor it for various firms depending on their unique data and requirements.

According to the study's findings, machine learning techniques can be used to forecast pay increase rates. The authors developed an ERP system-integrable machine learning-based system that forecasts wage rise rates. To determine the performance score, they created a formula and a scale that was unique to each study. The machine learning model may update itself to reflect changes in the business requirements thanks to frameworks created by the authors. Using web services created in accordance with the features of the study, the data specific to the study was obtained from task tracking systems. The authors come to the conclusion that the suggested system may be utilized by institutions and organizations to anticipate wage growth rates and that it can be built using a microservice architecture to accommodate future corporate requirements. The study's conclusions suggest that machine learning-based methods can be employed to forecast wage rise rates, which can assist businesses in making knowledgeable judgments regarding pay raises and enhance worker happiness and productivity.

The authors provide a thorough description of the proposed system architecture and the process used to design the system, which is one of the study's strong aspects. In addition, they offer a cutting-edge scale for figuring out the study-specific performance score. The features were generated by the authors using data from consulting firms, and they obtained research-specific data from task management programs utilizing web services created in accordance with the features in the study. The machine learning model

has the capability to autonomously update, adapting to changes in business requirements through frameworks developed by the authors. Finally, the authors propose that a micro-service architecture might be used to build the proposed system to accommodate future business requirements. Overall, the study offers a thorough method for estimating pay.

With a focus on AI and neural network models, this review of the literature explores the complex area of compensation prediction using employee data. It seeks to offer a thorough review of the state of the field's research, methodologies, difficulties, and new trends. We want to acquire insights into the methodology used, the parameters taken into account, the accuracy attained, and the practical consequences of AI-driven salary prediction by probing the corpus of existing research. We also want to identify gaps in the current research environment and suggest directions for future research that can progress the application of AI in HR practices, eventually benefiting businesses and their employees.

Another paper (Bansal, Narang, Sachdeva, Kashyap, & Panda, 2021) This paper is an empirical analysis of regression techniques for predicting house prices and salaries. The authors compare the performance of two regression techniques, decision tree and linear regression, using two datasets: employee salaries and house prices. They also discuss the importance of feature selection in machine learning prediction and the process of splitting datasets into training and test sets. The authors conclude that decision tree provides the best accuracy for predicting house prices, while linear regression is more suitable for predicting employee salaries. The findings of this paper have potential applications in machine learning and data analysis. The research question of this paper is to compare the performance of two regression techniques, decision tree and linear regression, for predicting house prices and salaries. This work presents an empirical analysis of prognostication and prediction using regression techniques. The objectives of this paper are to identify the importance of feature selection in machine learning prediction, to discuss the process of splitting datasets into training and test sets, and to determine which regression technique provides the best accuracy for predicting house prices and salaries.

The dataset used in this paper is an open-source dataset from KaggleInc. The dataset contains dependent and independent attributes or factors, which are called features in supervised machine learning algorithms.

The data collection method is not explicitly stated in the paper, but it can be inferred that the authors collected the dataset from KaggleInc, a platform for data science and machine learning competitions. The authors likely downloaded the dataset from KaggleInc and used it for their analysis.

The paper does not explicitly mention any difficulties faced by the authors during their research. However, it is common for researchers to face challenges during the data collection and analysis process, such as missing or incomplete data, data cleaning, and selecting appropriate features for the analysis. Additionally, the authors may have faced challenges in selecting the appropriate regression technique for each dataset and in interpreting the results of their analysis. However, the paper does not provide any specific information about the difficulties faced by the authors.

The results presented in this paper indicate that decision tree regression yields higher accuracy in predicting house prices, whereas simple linear regression performs better in predicting salaries. The authors also emphasize the significant impact of feature selection on machine learning predictions and underscore the importance of dividing the dataset into training and test sets to achieve accurate predictions.

The conclusions of this paper are that the regression technique selection depends on the specific dataset and the problem being addressed, and that feature selection and dataset splitting are important considerations for achieving accurate predictions. The authors also suggest that further research is needed to explore other regression techniques and to investigate the impact of different feature selection methods on prediction accuracy.

The implications of this paper are that machine learning techniques can be used to predict house prices and salaries with reasonable accuracy, and that decision tree and simple linear regression are effective techniques for these types of predictions. The paper also highlights the importance of careful feature selection and dataset splitting for achieving accurate predictions. These findings may be useful for researchers and

practitioners in the fields of real estate and human resources, as well as for anyone interested in using machine learning for prediction and prognostication.

The strong points of this paper include:

1. Clear research methodology: The paper provides a clear and well-defined research methodology, which includes data collection, dataset splitting, and the use of various regression techniques for prediction.

2. Use of open-source dataset: The authors used an open-source dataset from KaggleInc, which is a well-known platform for data science and machine learning competitions. This makes the results of the study more transparent and reproducible.

3. Comparison of different regression techniques: The authors compared the performance of several regression techniques, including decision tree, simple linear regression, SVM, Lasso regression, and logistic regression. This provides valuable insights into the strengths and weaknesses of each technique for predicting house prices and salaries.

4. Emphasis on feature selection: The authors emphasized the importance of feature selection in machine learning prediction and provided insights into the impact of different feature selection methods on prediction accuracy.

5. Implications for real-world applications: The findings of this paper have important implications for real-world applications, such as predicting house prices and salaries. The authors' recommendations for selecting appropriate regression techniques and performing feature selection can help practitioners achieve more accurate predictions.

Regarding the weak points of this paper, there are a few limitations that should be considered:

1. Limited dataset: The authors used a single dataset from KaggleInc, which may not be representative of all real estate and salary prediction scenarios. The dataset also has a limited number of features, which may not capture all the factors that affect house prices and salaries.

2. Lack of explanation of some techniques: The paper does not provide a detailed explanation of some of the regression techniques used, such as SVM and Lasso regression. This may make it difficult for readers who are not familiar with these techniques to understand the results.

3. Limited discussion of results: The paper provides a brief discussion of the results, but does not provide a detailed analysis of the strengths and weaknesses of each regression technique. This may limit the usefulness of the paper for practitioners who are interested in selecting the most appropriate technique for their specific problem.

4. Lack of comparison with other studies: The paper does not compare its results with other studies in the field, which may limit the generalizability of the findings.

5. Lack of consideration of other factors: The paper focuses on the impact of regression techniques and feature selection on prediction accuracy but does not consider other factors that may affect the accuracy of predictions, such as data quality, model complexity, and overfitting.

(Chakraborti, 2014) This paper offers a comparative study examining different classification algorithms for the prediction of salary classes among employees. The study uses the UCI census dataset and evaluates the performance of algorithms such as decision tree, Bayesian network, k-nearest neighbor, and support vector machine. The study concludes that decision tree and Bayesian network classifiers demonstrate superior performance in predicting salary classes. The paper also discusses the challenges faced by corporations in determining optimum salary offers for high-performance employees. The forthcoming research entails collecting real-life company data and conducting salary predictions using the algorithms outlined in the study.

The research question of this paper is to find out the best classification algorithm that can predict the salary class of employees based on a given set of attributes. The aims of this paper are to run various classification engines on the UCI census dataset, compare their prediction performance, assess the impact of feature selection techniques, and compare the training time of the classification engines. The goals of this paper include establishing a comprehensive understanding of the suitability of these algorithms for predicting employee salaries, assessing the performance metrics of diverse machine learning algorithms in addressing salary classification problems, and determining whether a subset of features can be effective for training, potentially achieving optimal performance instead of using the complete set. The data used in this paper is the UCI census dataset, which is a well-known dataset for machine-learning experiments. The dataset contains information about individuals such as age, education, occupation, marital status, and income. The data collection method for this dataset is not mentioned in the paper.

The paper discusses several difficulties faced in the salary prediction problem. One of the main difficulties is the unbalanced nature of the UCI census dataset, where 70% of the data belongs to the <=S0K salary category. This could lead to bias in prediction. Another difficulty is the avoidance of over-fitting in algorithmic models. The paper suggests using SMOTE filter of WEKA or random under-sampling to address this issue. Additionally, the paper discusses the challenge faced by corporations in determining optimum salary offers for high-performance employees.

The findings of this paper suggest that the Random Forest algorithm performs the best for salary classification problem, followed by the J48 decision tree algorithm. The paper also found that feature selection techniques can improve the performance of the classification algorithms. The conclusions drawn from this paper are that machine learning algorithms can be effectively used for salary prediction of employees, and that feature selection techniques can be used to improve the performance of the algorithms. The implications of this paper are that the findings can be used by companies to automate the salary prediction process and make enterprise compensation management more efficient. The paper also proposes that future research should encompass collecting real-life

company data to validate the findings of this experiment. Additionally, it suggests enhancing the UCI census dataset by incorporating performance attributes to create an experiment that closely mimics real-life salary prediction scenarios.

The strong points of this paper include:

1. The paper provides a comprehensive comparison of several machine learning algorithms for salary classification problems using the UCI census dataset.

2. The paper discusses The effect of feature selection strategies on the efficiency of the classification algorithms.

3. The paper provides detailed experimental results and performance metrics for each algorithm.

4. The paper suggests ways to address the difficulties faced in the salary prediction problem, such as the unbalanced nature of the dataset and the avoidance of over-fitting.

5. The paper provides implications for the use of machine learning algorithms in real-life salary prediction and compensation management.

potential weak points of this paper could include:

1. The study only uses one dataset (UCI census dataset) which may not be representative of all salary prediction problems.

2. The study does not consider the impact of other factors such as location, industry, and experience on salary prediction.

3. The study does not provide a detailed explanation of the algorithms used, which may make it difficult for readers to replicate the experiment.

4. The study does not discuss the ethical implications of using machine learning algorithms for salary prediction, such as potential bias and discrimination.

5. The study does not provide a comparison of the computational resources required for each algorithm, which may be important for practical applications.

In (Martin et al., 2018) presents a case study on salary prediction in the IT job market using data from a Spanish IT recruitment portal. The study analyzes 4,000 job offers and identifies key factors in determining salary ranges. The paper recommends using various semi-manual feature preprocessing techniques to improve data quality, formatting, and standardization of collected data. It also tries to lower dimensionality by a factor of 10 to improve prediction accuracy. The research presents the problem of predicting a person's salary as an example of classifying data. It compares in detail various approaches to clustering, from SVM and MLP down through random forests, AdaBoost, and their ensembles. The goal is to find the model that gives us the highest accuracy in predicting salary ranges. The paper concludes with an exhaustive assessment of classifiers, providing a detailed summary of the obtained results.

It also suggests the use of human preprocessing techniques on features to improve data quality, format, and standardize collected information. In addition, it aims to cut dimensionality by a whopping 10 times to increase prediction accuracy. This research gives the salary prediction problem a classification form and surveys in depth of various classifiers, such as SVMs, MLPs (Multi-Layer Perceptron's), random forests, and AdaBoosters. The aim is to find the model with the highest accuracy in predicting salary ranges. The paper is rounded out by a detailed assessment of classifiers to provide an outlined review of all research results. Using tree-based ensemble methods, the primary aim is to build a more accurate salary range classifier. Other goals include manual feature preprocessing techniques for data refinement, formatting, and standardization. Further, it tries to examine the aspects of a job post that have an important influence on salary and see exactly what their relationships are. The research also seeks to extract primary data-driven profiles generated through skill-set-based aggregation and makes a comprehensive comparison of several classifiers. Which model has the greatest accuracy in predicting salary ranges?

The information for this study was gathered from a Spanish IT recruitment portal, yielding a dataset of 3,970 job postings. To collect the necessary data from the website, a Python-based web crawler was developed. This crawler ran daily between December 2015 and April 2016, with duplicate entries subsequently eliminated. During the data cleaning

phase, posts with missing values were removed, and potential conflicts in the data format (such as text encoding issues) were resolved. In the manual feature engineering phase, irrelevant features were discarded, and others were standardized—leveraging domain knowledge, some features were converted into numerical representations.

A small sample, high dimensionality, and noise in the collected data proved to be serious obstacles for this study. However, predicting salaries was no easy task. Clear salary signals were posted for less than around 4,000 posts over five months on the website. Because the data was so high-dimensional, 200 different features were assigned to each post in vector form. Thirdly, noise in the data also made it difficult to extract useful information.

The study found that the proposed manual feature preprocessing method improved the prediction accuracy by reducing dimensionality by 10 times. The study also found that the most important areas in a job posting that have a greater effect on salary are job position, experience, and education level. the study identified the most important data-driven profiles that are most valuable and sought after in job offers within the IT job market. The study formulated the salary prediction problem as a classification task and compared several classifiers to determine which model best predicts the wage range in terms of accuracy. The study concluded that tree-based ensembles are the best classifiers for predicting salary ranges in the IT job market. According to a study, the proposed method might be effectively used by an e-Recruitment website. It can also make automatic classification of job posts by salary range, especially those in which the actual salary offered is not explicitly given. Apart from this, the method could also form a subsection in developing the job recommender system.

The strong points of the study include the proposed manual feature preprocessing method, which improved the prediction accuracy by reducing dimensionality by 10 times, and by framing the salary prediction problem as a classification task, the study achieved improved accuracy by concentrating on discrete salary ranges rather than continuous values. The research involved a comparison of multiple classifiers to identify the model with the highest accuracy in predicting salary ranges. Additionally, it revealed key data-

driven profiles obtained through skill-set-based aggregation. These findings hold practical utility, as they can be effectively applied by an e-Recruitment website for automated categorization of job posts by salary range, even in cases where the actual offered salary is unspecified. Furthermore, the results can serve as a foundational element in constructing a job recommender system.

This paper (Khongchai & Songmuang, 2016) presents a method of salary prediction that compares the profiles of current and graduated students using data mining to enable individual-based salary prediction. The system aims to motivate students during their studies by providing a positive outlook. A overview of prior studies on data mining methods used in the field of education is included in the study , an explanation of the proposed system, experimental results, and a conclusion with a summary of the findings. The results indicate that the K-NN technique gave the best accuracy at 84.69%, while the Multilayer perceptron returned the lowest accuracy at 38.08%. The system has received positive feedback from students, who recommend using it prior to the entrance exam to assist students in making logical decisions about their academic route and aim before selecting a faculty member and program. The research question of this paper is how to identify effective data mining methods for wage prediction in order to inspire students to study. The aim of the paper is to present a salary prediction system that uses data mining techniques to compare the profiles of graduated and current students in order to offer individual-based wage prediction.

The data used in this paper is the historical data of students who graduated from Rajamangala University of Technology Thanyaburi, Thailand, between 2006 and 2015. The data includes gender, student faculty, student program, type of work, experience of job training, received certification, total Grade Point Average (GPA), salary, address, telephone number, and e-mail. The data was collected from the university's database. The authors also conducted a questionnaire survey with a sample of users to evaluate the system's usability and effectiveness. However, the paper does not provide specific details about the questions asked in the survey.

The paper mentions several difficulties faced during the development of the salary prediction system. One of the main difficulties was the lack of data on students' salaries, which made it challenging to train and test the system accurately. Another difficulty was the lack of a standard format for the data, which required the authors to preprocess the data to ensure consistency and accuracy. Additionally, the authors faced challenges in selecting the most appropriate data mining techniques to use in the system, as there are many techniques available, and each has its strengths and weaknesses. Finally, the authors faced difficulties in evaluating the system's effectiveness, as there is no standard method for evaluating the accuracy of salary prediction systems.

The paper shows that the highest accuracy was 84.69% when the K-NN (k nearest neighbors) data mining technique was used, while the lowest figure of all is found with Multilayer Perceptron at only 38.088%. Thus, the conclusion is that the proposed system for predicting salaries can in fact help to make individualized estimates by comparing the profiles of current and completed students. Several other features include: The system can also help students select faculties and fields of study with good prospects for income in the job market. The strong points of the paper include:

1. The paper addresses an important issue of predicting salaries for students, which can motivate them during their studies and assist them in planning their studies successfully to achieve good salaries.

2. The paper proposes a salary prediction system using data mining techniques, which can be used to compare the profiles of graduated and current students in order to offer individual-based wage prediction.

3. The paper compares several data mining techniques and evaluates their accuracy in predicting salaries, providing insights into the strengths and weaknesses of each technique.

4. The paper presents experimental results using a large dataset of graduated student data, which adds credibility to the proposed system.

5. The paper provides positive feedback from students who used the system, indicating that the system can increase activity and motivation for studying and can be used as a guide for students to choose faculties and fields of study that will generate good incomes in the labor market.

6. The study proposes that the system can be used to teach parents about academic departments and programs, improving understanding of the goals of the research concerning parents' and students' professions and salaries.

Here are some weak points of the paper:

1. The paper acknowledges that the system has some limitations, such as the inability to adjust for changes in salaries, which can result in prediction errors.

2. The paper reports that some students provided negative feedback, such as the lack of clear identification of names/titles of careers or positions in the system, which can limit its usefulness.

3. The paper does not address any ethical concerns associated with the use of the proposed system, such as the potential for bias or discrimination in salary predictions based on certain attributes.

Another paper (Kablaoui & Salman, 2022) explores the benefits of data science and machine learning for predicting salaries in the USA. The authors collected a dataset of 20,000+ salaries and applied three supervised ML techniques (linear regression, random forest, and neural networks) to produce salary predictions. The outcomes indicate that neural networks surpass other machine learning models, achieving an accuracy level of 83.2%, whereas linear regression boasts the quickest training time at 0.363 seconds. In addition to presenting their results, the authors conducted a comprehensive literature review, comparing various machine learning models employed in salary prediction. It is crucial to note that the primary limitation of the paper lies in the specificity of the dataset used, implying that results may slightly differ for reduced categories within the dataset.

The research question of this paper is how to use data science and machine learning to predict salaries in the USA. The aim of the paper is to explore the benefits of these

techniques for salary prediction and to compare the performance of different ML models. The objectives of the paper are to gather a dataset of salaries in the USA, apply preprocessing techniques, prepare the training and testing data, and model the data using three supervised ML techniques (linear regression, random forest, and neural networks). The paper also aims to analyze and compare the output of the three models and to conduct a literature review to compare multiple ML models used for salary prediction.

The data used in this paper is a dataset of salaries in the USA. The dataset contains information on specific traits such as age, work sector, education, marital status, occupation, gender, hours per week, and country, as well as a column for the salary. The dataset aim is to predict whether an individual's annual salary exceeds $50,000. The data was collected from two sources: online research and an online survey. The researchers identified over 20,000 datasets related to salary prediction in the USA from Kaggle. To supplement this data, they administered an online survey to approximately 100 individuals in Kuwait, encompassing diverse backgrounds and featuring questions akin to those found in the dataset. Subsequently, the authors amalgamated all the gathered data into a single CSV file for further analysis.

The paper does not explicitly mention any difficulties faced during the research process. However, it is worth noting that the authors had to apply preprocessing techniques to the dataset to guarantee the data's quality and the valuable information it can extract. Additionally, all rows with Nan values had to be removed, which could affect their model prediction by having missing information. Additionally, the authors acknowledge that the obtained result is for only one type of dataset, so reduced categories of the dataset may slightly differ from the result.

The findings of this paper indicate that machine learning techniques can be used to predict salaries in the USA with high accuracy. The authors applied three supervised ML techniques (linear regression, random forest, and neural networks) to the dataset and found that the random forest model outperformed the other two models in terms of accuracy. The authors also conducted a literature review and compared multiple ML

models used for salary prediction, which helped them identify the strengths and limitations of each model.

The conclusions of the paper are that machine learning techniques can be used to predict salaries in the USA with high accuracy, and that the random forest is the most accurate model for this task. The authors also conclude that preprocessing techniques such as normalization, one-hot encoding, and binary encoding are necessary to ensure the quality of the data and the meaningful information it can extract.

The implications of this paper are that machine learning techniques can be used to predict salaries in the USA, which can be useful for individuals, companies, and policymakers. For individuals, salary prediction can help them negotiate better salaries or make informed decisions about their careers. For companies, salary prediction can help them make informed decisions about hiring, promotions, and compensation. For policymakers, salary prediction can help them identify and address issues related to income inequality and labor market dynamics.

Some of the strong points of this paper are:

1. The authors gathered an enormous collection of American wage data that includes columns for salaries as well as details on other characteristics like age, country, work sector, education, marital status, occupation, gender, and hours worked each week. The dataset's main objective is to forecast whether an individual's annual salary exceeds $50,000.

2. The authors applied three supervised ML techniques (linear regression, random forest, and neural networks) to the dataset and found that the random forest model outperformed the other two models in terms of accuracy.

3. The authors conducted a literature review and compared multiple ML models used for salary prediction, which helped them identify the strengths and limitations of each model.

4. The authors utilized preprocessing methods like normalization, one-hot encoding, and binary encoding to safeguard the data's quality and enhance its capacity to extract meaningful information.

5. The authors acknowledge the limitations of their study and suggest future research directions, such as combining different ML models to see how this could affect the overall accuracy of the ML algorithms.

However, some potential weak points of the paper could include:

1. The dataset used in the study only includes salaries in the USA, which may limit the generalizability of the results to other countries or regions.

2. The authors did not explore the interpretability of the models, which could be important for understanding the factors that contribute to salary prediction.

3. The authors did not discuss the potential ethical implications of using machine learning models for salary prediction, such as the risk of perpetuating bias or discrimination.

4. The authors note that as the result is based on a single type of dataset, the final result may differ slightly from the reduced categories of the dataset.

This paper (Khongchai & Songmuang, 2016) proposes a salary prediction system using a decision tree model with seven features to improve students' motivation in studying. Using 13,541 records of graduated student data in 10-fold cross-validation, the system was evaluated. The findings demonstrated that it successfully increased students' motivation and gave them an optimistic outlook on their future. The paper also includes suggestions from students for improving the system, such as adding missing features and using it before making entrance examination decisions.

The research question of this paper is how to improve students' motivation to study using a salary prediction system. The aim of the paper is to propose a decision tree model for creating a salary prediction system that can increase students' motivation to study. The objectives of the paper are to evaluate the system's efficiency using graduated student data,

to evaluate the system's effectiveness using questionnaires, and to gather feedback from students on how to improve the system.

The paper uses a dataset of 13,541 records of graduated student data to evaluate the efficiency of the salary prediction system. The data was collected electronically from the Office of Academic Resource and Information Technology at Rajamangala University of Technology Thanyaburi, Thailand. The data includes seven features, such as gender, major, and GPA, which were used in the decision tree model to predict the salary of students. The paper does not provide details on the data collection method, but it acknowledges the help of Puree, Taneth, Itikon, and everyone in AI Labs for collecting the electronic datasets.

However, it does mention some limitations of the study, such as the lack of data on some important features that could affect salary prediction, such as job position and company size. The paper also acknowledges that the dataset used in the study is limited to a single university in Thailand, which may not be representative of the entire population. Additionally, the paper notes that the system may not be applicable to other countries or regions due to differences in job markets and salary structures.

The paper's findings suggest the suggested salary prediction method may be able to address and solve the problems of previous systems by predicting salaries of individual students and presenting results in a way that does not require statistical expertise. The evaluation results of the salary prediction model indicate that the system has high precision, recall, and overall accuracy. The study also found that the system can increase students' motivation to study by providing them with a clear view of their future salary potential.

The paper concludes that the proposed salary prediction system can be an effective tool for increasing students' motivation to study and making informed decisions about their academic path. The system can also be improved by adding missing features and expanding the dataset to include more diverse samples. The implications of the study are that the proposed salary prediction system can be used by universities and educational institutions to help students make informed decisions about their academic path and career

choices. The system can also be used by employers to predict the salary potential of job candidates and make informed hiring decisions. The study also highlights the importance of using data-driven approaches to improve education and career outcomes.

The paper has several strong points, including:

1. Novelty: The proposed salary prediction system is a novel approach to increasing students' motivation to study and making informed decisions about their academic path and career choices.

2. Evaluation: The paper provides a detailed evaluation of the salary prediction model using precision, recall, and overall accuracy measurements, which demonstrates the effectiveness of the system.

3. User feedback: The paper includes feedback from users of the system, which provides insights into the system's usability and potential for increasing students' motivation to study.

4. Implications: The paper discusses the implications of the study for universities, educational institutions, and employers, which highlights the possible effect of the proposed system on education and career outcomes.

5. Clarity: The paper is well-organized and clearly presents the research problem, methodology, results, and conclusions, which makes it easy to understand and follow.

Chapter 3

Methodology

3.1 Research Design

The methodology employed in this research aims to develop an effective model for predicting salaries based on job titles and associated skills within the Turkish job market. To achieve this objective a multi-step approach was undertaken, encompassing data collection, preprocessing, and the application of various machine learning algorithms. The primary dataset manually collected from a prominent job search website in Turkey, consists of four key columns; job titles, average salary, skills, and top sector categorization, In the following sections we delve into the intricacies of our methodology, beginning with the transformation of textual data into numerical features using TFID vectorization. Subsequently, we explore the initial application of machine learning algorithms, Linear Regression, SVM, Decision Tree Regression, and Random Forest Regression for salary prediction. Acknowledging the initial performance limitations, the methodology evolves to incorporate KNN clustering, categorizing job sectors and salaries into three distinct clusters. This clustering approach, along with the introduction of a novel "skill score" feature derived from skill frequency analysis, significantly enhances the predictive capabilities of the model. The ensuing sections provide a detailed account of each methodological step, offering transparency into the choices made and their implications for the overall efficacy of salary prediction in the Turkish job market.

3.2 Data Collection

The Data was collected manually from one of the most prominent job search websites in Turkey, the website contains around 2800 job listing but unfortunately, most of them lack information about salaries or other features so we filtered out the jobs that have at least 100 participants who shared their salary information with the platform.
The dataset contains salary information for 117 job titles.

3.3 Data Preprocessing

The data was entered into an Excel spreadsheet. It consisted of four main columns, job titles, average salary, skills, and top sector with 117 rows. The data have been cleaned by removing unrelated skills, for example adobe photoshop skill is unrelated to a barista job.

Label encoding was used to transform the “top sectors” column into encoded numerical values. Label encoding is a simple technique for converting categorical variables into numerical labels. It is commonly employed to prepare categorical data for utilization in data analysis and machine learning algorithms that require numerical input. (Bishop, 2006) And by using the “encoded top sectors” column and the average salary column we clustered the data by K-Nearest Neighbors algorithm (KNN).

The Elbow method showed that the data can be clustered into 3 clusters.

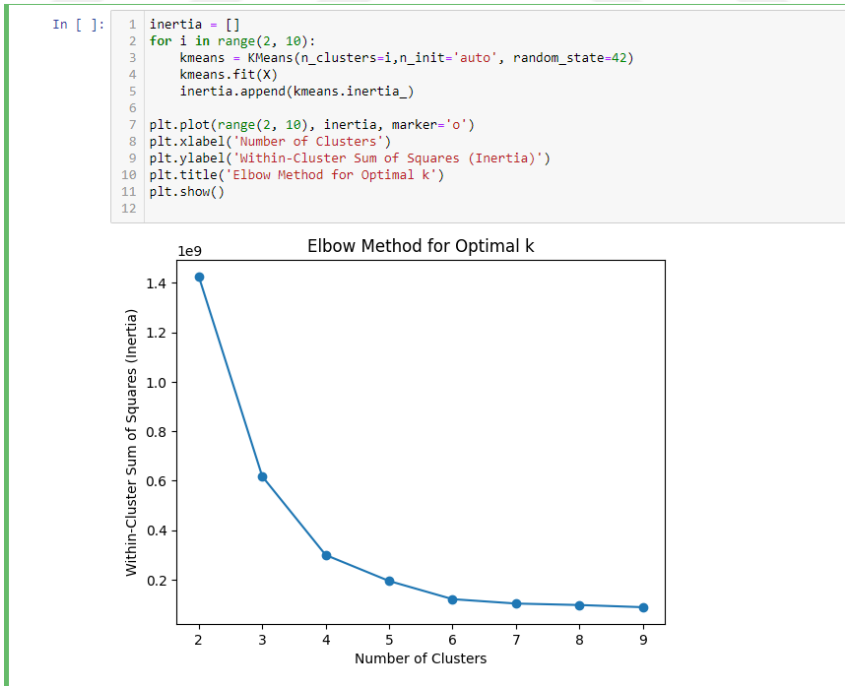

Figure 4. Elbow method for optimal k

we gave those clusters a number [0,1,2]

```
In [ ]:
# Plot clustered jobs
plt.figure(figsize=(12, 5))

# Loop through clusters and plot data points
for i  in range(n_clusters):
    cluster_data = df[df['cluster'] == i]
    plt.scatter(cluster_data['encoded_sectors'], cluster_data['average salary ( TL)'], label=f'Cluster {i}', marker='o')

plt.xlabel('Encoded sectors')
plt.ylabel('Average Salary (TL)')

plt.title('Clustered Sectors based on Average Salary')
plt.legend()
plt.show()
```

Figure. 5 Clustering result

As shown in Figure 5 clusters 0,1,2 represent high, medium, and low salary respectively. A new column generated in our dataset "cluster" is ready to be used in our machine learning prediction algorithm.

3.4 TF-IDF Vectorizer

Term frequency inverse document frequency or TF-IDF is gradually being adopted by machine learners as an effective way of transforming text into meaningful numeral representations which are crucial for prediction purposes. The following formula is used to calculate it:

$$\text{TF IDF}(\mathbf{t},\mathbf{d}) = \text{TF}(\mathbf{t},\mathbf{d}) * \text{IDF}(\mathbf{t})$$

$$\text{TF}(\mathbf{t},\mathbf{d}) = \left(\frac{freq(t,d)}{\sum_i^n \text{freq}_{(t_i,d)}}\right)$$

$$\text{IDF}(\mathbf{t}) = \log\left(\frac{N}{\text{count}(t)}\right)$$

Term Frequency (TF) is the measure used to evaluate the significance of a word or phrase in a singular text. Local density value is obtained as a ratio which results from

the term’s frequency of occurrence and the total number of words in the document. For instance, if the term "machine learning" manifests three times in a document containing 100 words, the corresponding TF value would be 0.03.

However, Inverse Document Frequency (IDF) serves as a weighting factor across all the documents of a collection. IDF is achieved through simply taking the logarithm on the number of documents with a word/phrase divided by the total number of documents.

The weight TF-IDF of a word or phrase is calculated by multiplying its TF value by its IDF value. This weighting represents the overall importance of a word or phrase for a particular paper in the collection.

TF-IDF vectorization is usually employed in natural language processing. It's used in tasks like categorizing text, summarizing text, and finding information. It's an uncomplicated yet powerful method that enhances machine learning algorithms working on text data.

TF-IDF vectorizer have been used on skills text in our dataset alongside with the encoded values of the job title and the cluster as inputs to machine learning algorithms [Linear regression, decision tree, random forest, XGBoost and support vector regression.]

3.4 Machine Learning Algorithms Used for Prediction

We have implemented five machine learning algorithms to predict salaries:

1. Linear regression

   Linear regression is a supervised machine learning algorithm designed to establish the linear relationship between a dependent variable and one or more independent variables. (Li, 2022)

2. Support Vector Machine (SVM)

   A supervised machine learning algorithm that can be used for both classification and regression tasks for purposes of regression, SVM seeks a

hyperplane which provides the optimal separation of the datapoints into various clusters. A hyperplane is selected to achieve the maximum margin (distance of the hyperplane to nearest clusters' data point). The support vector machine (SVM) is a very potent algorithm that can work with non-linear relationships of the independent variables and target's variable. Yet it is computationally more complex than linear regression. (Montesinos López et al., 2022).

3. Decision Tree Regression

Supervised machine learning technique through which a tree-like structure is built to predict a continuous dependent variable. It has several nodes and branches forming a tree. Each line indicates a decision and every subsequent is either of its outcomes. This will ensure that the data is broken down into increasingly smaller portions until each subset only has data points with the same value. A simple, quick, and interpretable decision tree regression algorithm. Though, this kind of model can suffer an unwanted phenomenon known as fitting, meaning that the learning curve for training data is too steep and the generalization capability in relation to new data decreases. (Nisbet et al., 2009)

4. Random Forest Regression

Supervised machine learning algorithm called random forest regression is an ensemble approach. The ensemble learning is an approach in which it combines several models to improve the overall accuracy of the model. Random forest regression constructs an aggregation of decision trees wherein every single tree has been taught with only unique sets of figures. Finally, all the trees give their prediction to determine the final prediction. Random Forest Regression is an important algorithm, which can be expected not to suffer from as much overfitting

as compared to Decision Tree Regressor. It tends to be more computationally demanding compared to the decision tree regression. (Breiman, 2001).

5. XGBoost

XGBoost is the short form of Extreme Gradient Boosting, which is a robust machine learning technique, established on the gradient boosting scheme. The ensemble learning approach, wherein several weak learners are combined to form one powerful learner referred to as gradient boosting. XGBoost can solve regression and classification problems very efficiently, which has been demonstrated in many applications.

One of the most important advantages of XGBoost is to handle complex relationships between independent variables and dependent ones. The algorithm is also capable of handling missing values and categorical data making it a multi-purpose algorithm fit for different types of data. Moreover, XGBoost is quite fast and scalable enough for big data sets. (Chen & Guestrin, 2016)

To assess the efficacy of these algorithms, the mean error value was selected as the performance metric. This metric provides a comprehensive measure of the average difference between predicted and actual salary values, offering valuable insights into the accuracy of the initial predictions generated by the machine learning models.

The dataset is divided into training and testing sets, with the training set comprising inputs along with their corresponding outputs. This division is essential for the model to learn from the training set and then apply the acquired knowledge to the testing set. The trained model endeavors to predict the outputs of the test set based on its learning from the training set.

We performed the train-test split mechanism on a local computer core i5-10300H CPU @ 2.5 GHz with python 3.11 using Scikit-learn (Sklearn) a library in python, the test size given is 0.1.

3.5 Model Refinement – skill score method

After measuring the mean absolute error, a new way to optimize our model performance has been introduced. The frequency of each skill across all jobs title has been counted and we assigned a score to each skill based on its frequency.

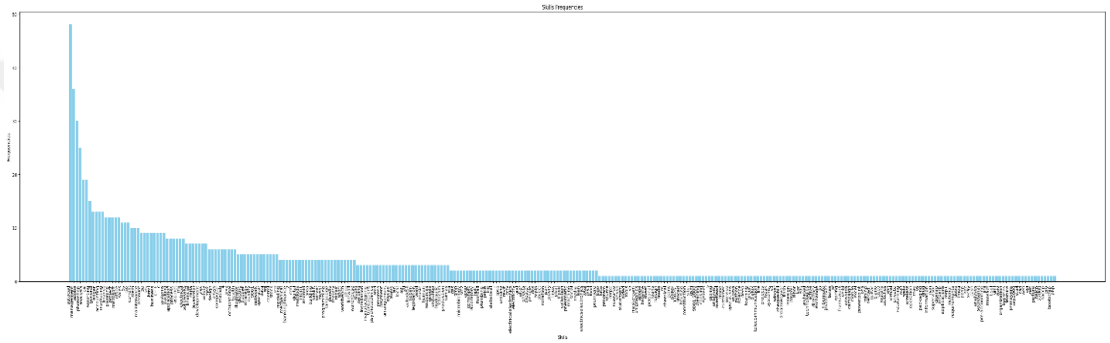

Figure 6. Skills frequency in the dataset

Therefore, we got a new column in our dataset describing the skill score for each job title. We have applied the previous five machine learning algorithms to predict the salaries again using skills score, job title and the cluster as inputs.

3.6 Limitations

There are some limitations regarding the quality of the data used for this study which require acknowledgement. In this regard, the data from the source showed some missing fields and discrepancies in the entries. Although there are some discrepancies, they may affect the final findings and need to be factored into while making conclusions. Also, the lack of information regarding people's professional experience gives rise to another major defect in our knowledge since experience normally represents a crucial factor in salary determination. More exhaustive datasets on experience need to be used in future studies that consider additional experience metrics for enhanced depth and accuracy

of salary forecasting.

However, it should be mentioned that this data only relates to the Turkish labor market, as it was gathered from Turkish job platform alone. As a result, these results could not fairly represent the changes in salaries across various geographic locations. Future iterations of this research may draw from larger databases obtained from a variety of labor markets, gaining a deeper understanding of potentially relevant pay predictor factors.

Chapter 4

Findings

This study investigated the effectiveness of two methods for predicting salary based on skills: this would involve using the skill score method as well as the TF-IDF method. Five machine learning algorithms were used with both methods: the models used in this case include linear regression, XGBoost, decision tree, SVR, and Random Forest. The performance of each algorithm was evaluated using Mean Absolute Percentage Error (MAPE). Generally, the skill score approach proved better at predicting salary than the TF-IDF technique. The probability of this lies in the fact that, unlike the skill score system, the TF-IDF algorithm doesn't account for the level of importance of certain skills.

Table.1

Mean Absolute Error and Mean Absolute Percentage Error Using TF-IDF Method

Algorithm	MAE	MAPE	R2
Linear Regression	2084.01	12.02	0.85
XGBoost	2767.47	13.26	0.73
Decision Tree	3643.33	16.37	0.48
SVR	1390.44	7.41	0.93
Random Forest	2855.63	13.47	0.73

SVR in the TF-IDF method demonstrated a mean absolute percentage error of 7.41%, showcasing a relatively lower degree of prediction error compared to other algorithms.

In order to understand the model better deferent scenarios were simulated by giving the model list of skills, fewer skills, and unrelated skills with fixed input values for the job title and the cluster.

Table.2

Predicted Salary with Deferent Skills Scenarios.

Algorithm	Predicted salary with full skills	Predicted salary with fewer skills	Predicted salary with unrelated skills
Linear Regression	16790.00	19026.04	21412.63
XGBoost	16790.03	18324.16	18817.84
Decision Tree	16790.00	16790.00	16790.00
SVR	21944.82	20763.56	20444.36
Random Forest	21944.82	20763.56	20444.36

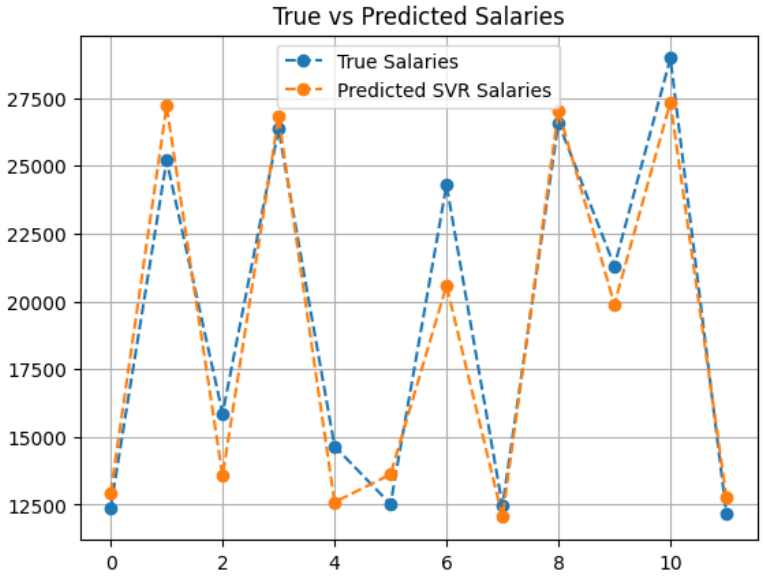

Figure.7 True Vs Predicted Salaries Using SVR Algorithms

Table 3 summarizes the mean absolute error for the five trained models after using skills score method.

Table.3
Mean Absolute Error and Mean Absolute Percentage Error for Skill Score Method

Algorithm	MAE	MAPE	R2
Linear Regression	1490.71	9.51	0.77
XGBoost	1655.55	10.66	0.76
Decision Tree	1392.50	8.25	0.84
SVR	1466.10	9.25	0.80
Random Forest	1024.02	6.55	0.87

Random Forest in the skill score method displayed the lowest mean absolute percentage error of 6.55%, reflecting its superior predictive accuracy compared to other algorithms in this method. And the Figure 10 shows the difference between predicted and true salary for random forest model.

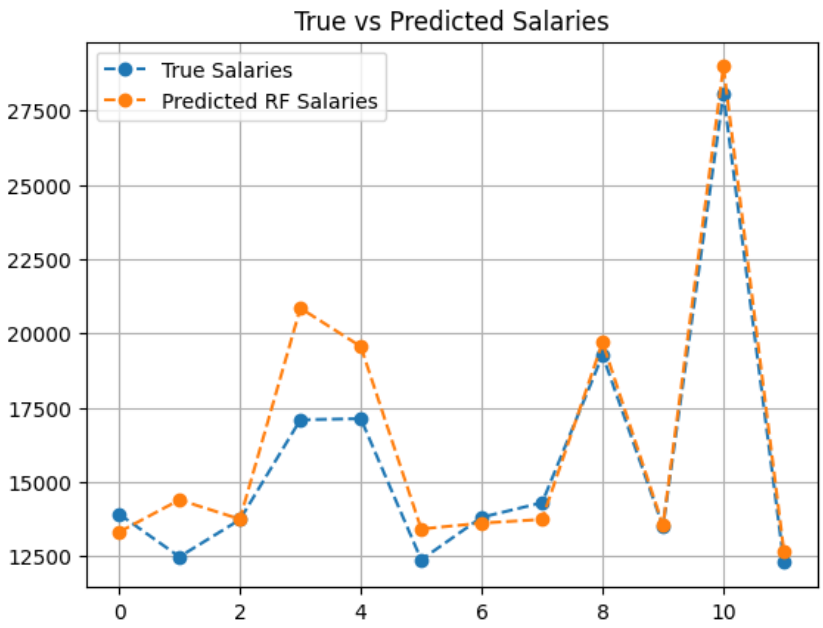

Figure.8 True Vs Predicted Salaries for random forest algorithm.

Chapter 5

Discussion and Conclusion

5.1 Discussion of Findings

Although the TF-IDF approach is frequently employed with text data, it poses some difficulties when predicting salaries. The different MAE values for each algorithm suggested that it was not suitable in the case of non-linear relationships, especially in terms of handling outliers. The random forest was tough, but its shortcomings were obvious. The above findings therefore call for adopting other approaches regarding feature representations for salary prediction activities. Regarding the inference test the tendency of linear regression to overpredict salaries, particularly in situations where skills become less relevant, stands out. Although the model's linearity may have an impact on its predictions, SVR's better performance, as evidenced by the higher R2 value, highlights the fact that sensitivity to skill relevance does not always equate to increased error. The R2 values indicate that SVR captures a significant portion of the variability in salary data. On the other hand, Decision Tree demonstrated stability in its prediction across several scenarios, this stability might come at the cost of capturing nuanced relationships between skills and salaries. Random Forest and XGBoost also showed stable predictions with minor variations, suggesting their adaptability to varying skill relevance.

Notably, as seen by the smaller mean absolute error, the Random Forest algorithm showed higher prediction accuracy when combined with the skill score technique. This highlights how important feature engineering is to improving machine learning models' capability for prediction. The predictions were clearly impacted when a 'cluster' feature was added to the data. The projected salaries for jobs in the same cluster were similar, indicating that the model gave this categorical variable a lot of weight. Further exploration into the cluster definitions and their impact on salary predictions could provide deeper insights into the underlying dynamics of the dataset.

The higher R2 values, especially with SVR and Random Forest, affirm the models'

ability to explain a significant portion of the salary variance based on the features used. The Random Forest algorithm demonstrated remarkable accuracy with a lower mean absolute error and a high R2 value, indicating its effectiveness in capturing the underlying patterns in the skill score method.

5.2 Conclusion

This study has assessed the efficacy of different machine learning models in predicting salary of job holder. The study employed two methods for representing skills: skill score method and TF-IDF. Five machine learning algorithms were evaluated using both methods: linear regression, XGBoost, decision tree, SVR, and random forest. The mean absolute percentage error (MAPE) was used in evaluating every single algorithm's performance.

The findings of the study suggest that the skill score method is generally more effective for predicting salary than the TF-IDF method. The results also showed that the performance of the algorithms varied depending on the method used. SVR outperformed other algorithms with TF-IDF method while random forest yielded superior results with skill score approach. in conclusion, this study conducted an investigation on salary prediction in the Turkish job market utilizing diverse machine learning methods and techniques. This investigation utilized TF-IDF vectorization, skill scores, and an analysis of various skill scenarios' effects on model predictions.

This research one of the first studies to thoroughly explore salary prediction using a leading job platform data in Turkish employment market, thereby creating a benchmark for best practice in the field. It sets the basis for future studies and promotes the application of advanced machine learning methods in salary prediction studies specific to Turkey.

Overall, this research is important due to its relevance to the Turkish job market, which bears on academic knowledge and stakeholders' practical solutions. By employing machine learning techniques and utilizing data from a popular job portal, this study

facilitates improvements in salary prediction approaches that can be used to enhance hiring practices and employment outcomes in Turkey. The Dataset generated presents a valuable resource with significant potential for researchers, academic institutions, and policymakers.

5.3 Recommendations for Future Research

Future work could focus on refining clustering techniques, exploring additional features, and assessing model generalization to diverse scenarios. Expanding the dataset from multiple job platforms to overcome the limitation of data quantity therefore explore the application of deep learning techniques such as neural networks that need bigger dataset.

Or explore the feasibility of providing inflation-adjusted salary predictions. This can offer more realistic and meaningful insights for individuals and organizations by accounting for the eroding effects of inflation on purchasing power.

REFERENCES

Bansal, U., Narang, A., Sachdeva, A., Kashyap, I., & Panda, S. P. (2021). Empirical analysis of regression techniques by house price and salary prediction. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012110). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012110/pdf

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Chakraborti, S. (2014). A Comparative Study of Performances of Various Classification Algorithms for Predicting Salary Classes of Employees. International Journal of Computer Science and Information Technologies, 5(2). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.7996&rep=rep1&type=pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.

Das, S., Barik, R., & Mukherjee, A. (2020). Salary prediction using regression techniques. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19). 1

Dutta, S., Halder, A., & Dasgupta, K. (2018, November). Design of a novel prediction engine for predicting suitable salary for a job. In 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 275-279). IEEE.

Eichinger, F., & Mayer, M.. (2022). *Predicting Salaries with Random-Forest Regression*. Springer eBooks. https://doi.org/10.1007/978-3-031-18483-3_1

Goerke, Laszlo and Neugart, Michael, Thorstein Veblen, Joan Robinson, and George Stigler (Probably) Never Met: Social Preferences, Monopsony, and Government Intervention (January 23, 2020).

Görmez, Y., Arslan, H., Suat, S. A. R. I., & DANIŞ, M. SALDA-ML: Machine Learning Based System Design to Predict Salary In-crease. Advances in

Artificial Intelligence Research, 2(1), 15-19. https://dergipark.org.tr/en/download/article-file/2104668

Gurpreet, M., Bistania, K., Mahendroo, S., Dogra, N., & Tuteja, P.. (2023). *Salary Prediction System using Machine Learning*. *07*(05). https://doi.org/10.55041/ijsrem22822

Huynh, T. V., Nguyen, K. V., Nguyen, N. L.-T., & Nguyen, A. G.-T. (2019). Job prediction: From deep neural network models to applications. arXiv preprint arXiv:1912.12214.

Jafari, A., Rouhanizadeh, B., Kermanshachi, S., & Murrieum, M. (2020). Predictive Analytics Approach to Evaluate Wage Inequality in Engineering Organizations. *Journal of Management in Engineering*, *36*, 04020072.

Kablaoui, R., & Salman, A. (2022). Machine Learning Models for Salary Prediction Dataset using Python. 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA), no. https://doi.org/10.1109/icecta57148.2022.9990316

L. Breiman. Random forests. Machine Learning, 45(1):5–32, 2001. 18

Li, Haksun. (2022). *Linear Regression. Numerical Method Using Java,* Berkeley, CA: Apress.

Mallapaty, S. (2018). Predicting scientific success. Nature, 561(7723), S32-S32.

Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018). Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study. International Journal of Computational Intelligence Systems, 11(1), 1192-1209.

Montesinos López, O. A., Montesinos López, A., & Crossa, J.. (2022). *Support Vector Machines and Support Vector Regression*. https://doi.org/10.1007/978-3-030-89010-0_9

Nisbet, R., Elder, J. and Miner, G. (2009), *Handbook of Statistical Analysis and Data Mining Applications*, Burlington: Elsevier, ISBN: 978-0-12-374765-5

P. Khongchai and P. Songmuang, "Implement of salary prediction system to improve student motivation using data mining technique," *2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, Yogyakarta, Indonesia, 2016, pp. 1-6, doi: 10.1109/KICSS.2016.7951419.

P. Khongchai and P. Songmuang, "Improving students' motivation to study using salary prediction system," 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 2016, pp. 1-6, doi: 10.1109/JCSSE.2016.7748896.

Quan, T. Z., & Raheem, M. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature. Journal of Applied Technology and Innovation (e-ISSN: 2600-7304), 6(3), 70. https://www.researchgate.net/profile/MafasRaheem/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_A_Literature_Review/links/62e1236e4246456b55edfe48/Salary-Prediction-in-Data-Science-Field- Using-Specialized-Skills-and-Job-Benefits-A-Literature-Review.pdf

Ş. Demir İnan Özer, B. Ülke, F. Serhan Daniş and G. Keziban Orman, "Salary Prediction via Sectoral Features in Turkey," 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Biarritz, France, 2022, pp. 1-6, doi: 10.1109/INISTA55318.2022.9894130.

Saito, S., Nakano, T., Chiang, C. W., Kobayashi, T., Savage, S., & Bigham, J. P. (2019). TurkScanner: Predicting the hourly wage of microtasks. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019* (pp. 3187-3193). (The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019). Association for Computing Machinery, Inc. https://doi.org/10.1145/3308558.3313716

Şentürk, İ. (2015) Determinants of Expected Wages of Unemployed Workers in Turkey. *Modern Economy*, **6**, 808-815. doi: 10.4236/me.2015.67076.

Tansel, A., Dalgıç, B. & Güven, A. *Wage Inequality and Wage Mobility in Turkey*. *Soc Indic Res* **142**, 107–129 (2019). https://doi.org/10.1007/s11205-018-1908-1

Utilization of Multinomial Naive Bayes Algorithm and Term Frequency Inverse Document Frequency (TF-IDF Vectorizer) in Checking the Credibility of News Tweet in the Philippines. (2023). Utilization of Multinomial Naive Bayes Algorithm and Term Frequency Inverse Document Frequency (TF-IDF Vectorizer) in Checking the Credibility of News Tweet in the Philippines. abs/2306.00018. https://doi.org/10.48550/arXiv.2306.00018]

Viroonluecha, P., & Kaewkiriya, T. (2018). Salary predictor system for Thailand labour workforce using deep learning. 2018 18th International Symposium on Communications and Information Technologies (ISCIT). [9]

Wu, M., & Shu, S. (2018). Top management salary, stock ratio and firm performance: A comparative study of state-owned and private listed companies in China. Journal of Accounting in Emerging Economies. [10]

Zhang, Y., Li, J., Wang, Y., & Wang, J. (2022). Employee salaries analysis and prediction with machine learning. 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA).