शोध पत्र / Research Paper

क्लस्टर आकार और डेटासेट में परिवर्तन करके बायोडाटा वर्गीकरण पर के–मीन्स और रैंडम फॉरेस्ट का प्रभाव विश्लेषण

K-Means and Random Forest Impact Analysis on Resume Classification by Altering Cluster Size and Dataset

स्विटी महेश पाटील[1], डॉ. विनायक डी. शिंदे[2]

Sweety Mahesh Patil[1], Dr. Vinayak D. Shinde[2]

[1]Shree L. R. Tiwari College of Engineering, Thane, India

[2]Associate Professor, Shree L.R. Tiwari College of Engineering, Thane, India -401107

[1]sweetypatil6566@gmail.com, [2]vdshinde@gmail.com

https://doie.org/10.0524/VP.2024275063

सारांश

ऑनलाइन नौकरी भर्ती के इस युग में नौकरी चाहने वालों और भर्तीकर्ताओं को सटीक नौकरी वर्गीकरण की आवश्यकता है। भर्ती प्रक्रिया के कई चरणों का सबसे महत्वपूर्ण हिस्सा बायोडाटा को वर्गीकृत करना है जो कागजी कार्य और मानवीय त्रुटियों को बचा सकता है, और सीमित कर सकता है। संगठन की व्यावसायिक सफलता पर विचार करते समय उम्मीदवार का व्यक्तित्व सबसे महत्वपूर्ण तत्वों में से एक है। इस शोध पत्र में, के–मीन्स क्लस्टरिंग मशीन लर्निंग एल्गोरिथ्म (K-Means Clustering Machine Learning Algorithm) का उपयोग करके व्यक्तित्व का पता लगाया जा सकता है। जब उपयुक्त व्यक्तित्व का पता चल जाएगा तो बायोडाटा पीडीएफ (PDF) प्रारूप में इनपुट के रूप में सिस्टम को भेज दिया जाएगा। रैंडम फॉरेस्ट (Random Forest) वर्गीकरण एल्गोरिथ्म का उपयोग करके बायोडाटा को नौकरी के शीर्षक से मेल खाने वाले प्रतिशत के आधार पर वर्गीकृत किया जाएगा। इस प्रक्रिया के परिणामस्वरूप आवेदकों का तत्काल चयन हो जाता है। रेज़्यूमे (Resume) वर्गीकरण के अलावा यह शोध पत्र रेज़्यूमे वर्गीकरण पर के–मीन्स और रैंडम फॉरेस्ट एल्गोरिथ्म के डेटा सेट के अलग–अलग क्लस्टर आकार, और आकार के प्रभाव का विश्लेषण करने में भी मदद करता है। के–मीन्स और रैंडम फॉरेस्ट एल्गोरिथ्म पर विभिन्न क्लस्टर आकारों और डेटा सेटों के परीक्षण के प्रभाव के परिणामस्वरूप सटीक रेज़्यूमे वर्गीकरण का इष्टतम विन्यास होता है, जो भर्ती प्रक्रिया में सुधार करता है। इस प्रयोगात्मक अध्ययन का परिणाम 70% सटीकता दर्शाता है और एन्सेबल लर्निंग (Ensemble Learning) तकनीक या सपोर्ट वेक्टर मशीन (एसवीएम) एक पर्यवेक्षित मशीन लर्निंग एल्गोरिथ्म का उपयोग करके इसमें और सुधार करना संभव है।

Abstract

Job seekers and recruiters need accurate job classifications in this era of online job recruitment. The most critical part of the numerous steps recruiting process is classifying resumes that can save and limit paperwork and human errors. The candidate's personality is one of the most crucial elements when considering the professional success of the organization. In this paper, personality can be detected using K-Means Clustering Machine Learning (ML) Algorithm. When an appropriate personality is detected resume will be passed to the system as input in PDF format. Using Random Forest Classification Algorithm resume is classified based on percentage matched to the job title. This process results in the immediate selection of applicants. In addition to the

resume classification this paper will also help to analyze the impact of varying cluster size and size of data sets of K-Means and Random Forest (RF) Algorithms on resume classification. The impact of testing various cluster sizes and data sets on K-Means and Random Forest Algorithm results in the optimal configuration of accurate resume classification, which improves the recruitment process. This experimental study, results shows 70% accuracy and possible to improve further by using ensemble learning technique or Linear Support Vector Classifier (SVC).

मुख्य शब्द : बायोडाटा, बायोडाटा वर्गीकरण, नौकरी चाहने वाले, भर्तीकर्ता, वर्गीकरण, क्लस्टरिंग, मशीन लर्निंग (एमएल), के–मीन्स, रैंडम फॉरेस्ट (आरएफ), डेटा सेट।

Keywords: Resume, Resume classification, job seekers, recruiters, Classification, Clustering, Machine Learning, K-Means, Random Forest, data sets.

परिचय

बायोडाटा एक ऐसा दस्तावेज है जो कर्मचारी की शैक्षणिक योग्यता, कार्य अनुभव, क्षमताओं और उपलब्धियों के बारे में संक्षिप्त जानकारी देता है। नौकरी चाहने वाला व्यक्ति, नौकरी के नए अवसरों की खोज करता है। दूसरी ओर, भर्तीकर्ता संगठन में नौकरी की रिक्तियों को भरने के लिए जिम्मेदार होता है, जो उम्मीदवार के अपेक्षित कौशल की खोज करता है। नौकरी खोजने की प्रक्रिया में बायोडाटा का मुख्य रूप से उपयोग किया जाता है। नियोक्ता को नौकरी चाहने वालों की क्षमता का परिचय देने के लिए बायोडाटा आवश्यक है। कुशलतापूर्वक संरचित बायोडाटा उम्मीदवार को भीड़ में अलग दिखने में मदद करता है। थोक में भर्ती करते समय, संगठन को कई संसाधनों की आवश्यकता होती है। भर्ती की पूरी प्रक्रिया समय लेने वाली साबित हो रही है। नियोक्ताओं को इंटरव्यू आयोजित करने के लिए विभिन्न स्थानों की यात्रा करनी पड़ती है, और इंटरव्यू देने वाले आवेदकों की विशेषताओं को याद रखना असंभव है। इसका समाधान बायोडाटा वर्गीकरण है [1]। बायोडाटा वर्गीकरण का उपयोग करते हुए, बायोडाटा को कुछ विशेषताओं के आधार पर वर्गीकृत किया जाता है, जो चयन प्रक्रिया को व्यवस्थित करता है। बायोडाटा का वर्गीकरण ML तकनीक का उपयोग करके किया जाता है। ML आर्टिफिशियल इंटेलिजेंस (AI) का एक हिस्सा है जो चयन प्रक्रिया को सुव्यवस्थित करने में मदद करता है। AI एक ऐसी तकनीक है जो स्मार्ट मशीनें बनाती है जो इंसान की तरह काम करती है। इसे हल करने के लिए शैक्षिक योग्यता, कार्य अनुभव, क्षमताओं और उपलब्धियों के अनुसार स्वचालित रूप से बायोडाटा को स्क्रीन करने के लिए विभिन्न मशीन लर्निंग एल्गोरिथ्म का उपयोग किया जाता है। यह स्वतंत्र रूप से सत्यापन करके पक्षपात को कम करने में मदद करता है। वर्गीकरण और क्लस्टरिंग, ML में दो महत्वपूर्ण तकनीकें हैं [2]। वर्गीकरण सुपरवाइज्ड लर्निंग की तकनीक है जबकि क्लस्टरिंग अनसुपरवाइज्ड लर्निंग की तकनीक है। वर्गीकरण लेबल किए गए डेटा सेट पर आधारित है जहां आउटपुट श्रेणी ज्ञात है। एक डेटा सेट में विभिन्न प्रकार के डेटा शामिल होते हैं जिनमें टेक्स्ट या छवि, संख्यात्मक डेटा आदि शामिल होते हैं जो आकार के अनुसार भिन्न होते हैं। क्लस्टरिंग का उद्देश्य डेटा में छिपे हुए पैटर्न या संरचनाओं का पता लगाना है। इस शोध पत्र का उपयोग प्रशिक्षण डेटा सेट की पहचान करने, उसी डेटा को मॉडल में पास करने के लिए किया जाता है जो के–मीन्स और रैंडम फॉरेस्ट जैसे मशीन लर्निंग एल्गोरिथ्म को लागू करके डेटा को उचित रूप से वर्गीकृत करता है। इसे प्राप्त करने के लिए, क्लस्टर आकार, प्रशिक्षण और परीक्षण डेटा सेट और विभिन्न मशीन लर्निंग एल्गोरिथ्म जैसे विभिन्न कारकों की पहचान करने के लिए कठोर साहित्य अध्ययन किया जाता है। इस प्राथमिक शोध में इस बात पर विचार किया गया कि कैसे एचआर, एडवोकेट,

आर्ट्स, वेब डिजाइनिंग, मैकेनिकल इंजीनियर, सेल्स, डेटा साइंटिस्ट आदि नौकरी शीर्षक श्रेणियों का उपयोग विश्लेषण और बायोडाटा को उपयुक्त श्रेणी में परीक्षण करने के लिए किया जाता है। इसके अलावा अतिरिक्त कार्य भूमिकाओं पर विश्लेषण करना संभव है। शोध पत्र में निम्नलिखित शीर्षक शामिल हैं: दूसरे शीर्षक में साहित्य सर्वेक्षण की व्याख्या की गई है। तीसरे शीर्षक में प्रस्तावित प्रणाली और कार्यप्रणाली को सिस्टम आर्किटेक्चर के साथ समझाया गया है। चौथे शीर्षक में अलग–अलग क्लस्टर आकार और डेटासेट द्वारा बायोडाटा वर्गीकरण पर के–मीन्स और रैंडम फॉरेस्ट के प्रभाव विश्लेषण के परिणाम पर चर्चा की गई है।

साहित्य सर्वेक्षण

[3] में, रिज्यूमे की स्क्रीनिंग के लिए एक वेब एप्लिकेशन जो 220 रिज्यूमे का उपयोग करता है – 200 प्रशिक्षण के लिए और 20 परीक्षण के लिए। आवेदक–पक्ष, सर्वर साइड और भर्तीकर्ता–पक्ष वेब एप्लिकेशन के घटक हैं। आवेदक अपने बायोडाटा को आवेदक पक्ष पर स्रोत करते हैं, जिसे SpaCy और NLP फ्रेमवर्क का उपयोग करके प्राकृतिक भाषा प्रसंस्करण (NLP) पाइपलाइन द्वारा सर्वर साइड पर संसाधित किया जाएगा। भर्तीकर्ता, पद के लिए सर्वश्रेष्ठ उम्मीदवार का चयन करने के लिए बायोडाटा रैंक सूची प्रदर्शित की जाती है और स्कोर कैलकुलेटर द्वारा निर्देशित की जाती है। [4] में, इंटेलीजेंट डेटा प्रोसेसिंग के लिए एक मशीन–लर्निंग तकनीक का उपयोग संगठनात्मक उद्देश्यों के लिए किया है। लीनियर डिस्क्रिमिनट एनालिसिस, नाइव बेयस (NB), सपोर्ट वेक्टर मशीन (SVM), और के–नियरेस्ट–नेबर्स (KNN), लॉजिस्टिक रिग्रेशन (LR), RF, एडाबूस्ट (AdaBoost) और डिसीजन ट्री (DT) विधियों का उपयोग किया है। यह शोध पत्र विविध संगठनात्मक रणनीतियों के बारे में निश्चित जानकारी प्रदान करता है। [5] में, प्रारंभिक एक उम्मीदवार द्वारा अपना बायोडाटा अपलोड करना है। इस दस्तावेज को संपादन की आवश्यकता है क्योंकि यह अधूरा और अव्यवस्थित है। कई मॉड्यूल डेटा मॉड्यूल, मॉडल प्रशिक्षण मॉड्यूल और परीक्षण मॉड्यूल के रूप में पेश किए गए थे। इंटरनेट से लिंक्डइन प्रोफाइल यूआरएल (URL) का उपयोग करके सेलेनियम के माध्यम से उन्हें आगे बढ़ाना है, जो प्रासंगिक फील्ड की पहचान करेगा और डेटा रिकॉर्ड करेगा। इस प्रकार डेटा मॉड्यूल काम करता है। [6] में, यह प्रणाली तीन चरणों पर आधारित है। पहला चरण पूर्वानुमान लगाना है। मूल्यांकन मैट्रिक्स का उपयोग करके उचित वर्गीकरण एल्गोरिथ्म चुनें, जो भविष्यवाणी करता है कि पंक्ति शीर्षक है या नहीं। दूसरा चरण खंड निष्कर्षण है, अगला शीर्षक प्रकट होने तक सभी जानकारी निकालता है। अंतिम चरण खंड वर्गीकरण है जो बायोडाटा में शामिल कौशल के आधार पर विवरण करता है। [7] में, NLP और ML तकनीकों का उपयोग भर्ती करने वालों को उचित प्रोजेक्ट सौंपने के लिए किया है। प्रशिक्षण में नामांकित इकाई पहचान (NER) दृष्टिकोण और LR और KNN कैटलॉगिंग जैसे कई सॉर्टिंग मॉडल पेश किए है। [2] में, निष्कर्ष प्राप्त करने के लिए DT, RF, गॉसियन नाइव बेयस (GNB), और KNN का उपयोग किया है। एल्गोरिथ्म को सत्यापित करने के लिए परिशुद्धता, सटीकता और रिकॉल के तीन उपायों का उपयोग किया है। KNN की सटीकता 93% है, RF की सटीकता 95% है, DT की सटीकता 96% है, और GNB की सटीकता 99% है। [1] में, बायोडाटा पीडीएफ फार्मेट में पास किया है। ऑप्टिकल कैरेक्टर रिकग्निशन (OCR) तकनीक का उपयोग करके बायोडाटा से टेक्स्ट निकाला है। यूजर डिफाइंड क्लीनिंग फंक्शन का उपयोग करके साफ और सादा पाठ प्राप्त किया जाता है। बायोडाटा का वर्ग प्रशिक्षित SVM क्लासिफायर के माध्यम से प्राप्त किया जाता है। अंतिम परिणाम प्राप्त करने के लिए बायोडाटा को समराइजर में पास किया जो टोकनाइजेशन पर आधारित है। [8] में, पहला मॉड्यूल रेज़्यूमे पार्सर है जो रेज़्यूमे से महत्वपूर्ण जानकारी निकालने के लिए NLP का उपयोग

करता है। दूसरा मॉड्यूल स्वचालित प्रश्न और उत्तर जनरेटर (AQG) है। इस मॉड्यूल का उपयोग करके ऑन्टोलॉजी यानी SE, QA, BA, सिस्टम इंजीनियर, नेटवर्क इंजीनियर उत्पन्न किया जाता है, फिर सिस्टम अगले चरण में Q&A तैयार करता है। अगले मॉड्यूल में सिमिलैरिटी कैल्कुलेशन, चेक कीवर्ड और कॉन्फिडेन्स क्लासिफिकेशन है और आउटपुट उत्पन्न होता है। [9] में, भावनाओं को निर्धारित करने के लिए सिमेंटिक विश्लेषण का उपयोग किया जाता है। स्ट्रिंग खोज की मदद से, ट्वीट निकाले जाते हैं और फिर उन्हें RF, SVM और NB का उपयोग करके तीन अलग–अलग श्रेणियों यानी सकारात्मक, नकारात्मक और तटस्थ में भावना विश्लेषण के लिए प्रस्तुत किया जाता है। साथ ही, यह शोध पत्र RF और SVM की सटीकता का अनुमान लगाता है। इसके अतिरिक्त, ट्वीट्स की संख्या बढ़ाकर RF, SVM और NB की सटीकता का अनुमान लगाया जाता है। मशीन लर्निंग एल्गोरिथ्म के पैरामीटर जैसे डेटासेट और क्लस्टर आकार पहले से तय होते हैं और यह केवल एक आउटपुट उत्पन्न करता है। इस शोध पत्र का मुख्य उद्देश्य पैरामीटर के मूल्यों में भिन्नता के आधार पर बायोडाटा वर्गीकरण पर विश्लेषण करना है।

प्रस्तावित प्रणाली और कार्यप्रणाली

इस शोध पत्र का उद्देश्य, व्यक्तित्व पहचान के साथ–साथ बायोडाटा वर्गीकरण का उपयोग करके नौकरी के लिए उपयुक्त उम्मीदवार ढूंढना है। व्यक्ति की पहचान में व्यक्तित्व अहम भूमिका निभाता है। पर्सनालिटी डिटेक्शन का उपयोग करके, कोई यह पता लगाने में सक्षम हो सकता है कि व्यक्ति नौकरी की भूमिका के लिए उत्तरदायी है या नहीं। पर्सनालिटी मॉड्यूल में OCEAN मॉडल का उपयोग किया जाता है जो पांच डाइमेंशन्स पर आधारित है– कर्तव्यनिष्ठा (Conscientiousness-C), खुलापन (Openness-O), सहमतता (Agreeableness-A), बहिर्मुखता (Extroversion-E), और न्यूरोटिसिज्म (Neuroticism-N)। इस मॉडल का उपयोग व्यक्ति के पर्सनालिटी का पता लगाने के लिए किया जाता है। [10] उदाहरण के लिए, डाइमेंशन्स (O) वाले व्यक्ति अक्सर फ्लेक्सिबल, अनुकूलनीय, खुले दिमाग वाला, इंटरैक्टिव, आदि। इस प्रणाली के विकास के लिए, उपलब्ध विभिन्न पायथन के साथ कार्यक्रम को समृद्ध करें। प्रायोगिक सेटअप के उद्देश्य से पायथन के संस्करण 3.11.3 का उपयोग किया गया है, साथ ही सिस्टम प्रोसेसर i5, 7th जनरेशन और 6GB रैम का उपयोग किया गया है। निम्नलिखित चित्र–1 प्रस्तावित सिस्टम फ्रेमवर्क दिखाता है, जिसमें दो मॉड्यूल शामिल हैं:

I) पर्सनालिटी का पता लगाना जिसमें निम्नलिखित चरण शामिल हैं:
   a) पर्सनालिटी इनफार्मेशन गैदरिंग
   b) डेटा प्री–प्रोसेसिंग
   c) डेटा संवर्धन
   d) पैटर्न पहचान
   e) पर्सनालिटी लक्षण की पहचान
   f) पर्सनालिटी प्रेडिक्शन

II) बायोडाटा वर्गीकरण में निम्नलिखित चरण शा. मिल हैं:
   a) डेटा एकत्रीकरण
   b) प्रशिक्षण डेटासेट पर टेक्स्ट वेक्टराइजेशन
   c) नौकरी शीर्षक लेबल एनकोडर
   d) pdf प्रारूप में इनपुट बायोडाटा
   e) डेटा प्री–प्रोसेसिंग
   f) परीक्षण डेटासेट पर टेक्स्ट वेक्टराइजेशन
   g) नौकरी शीर्षक भविष्यवाणी.

निम्नलिखित चरण प्रस्तावित सिस्टम फ्रेमवर्क को दर्शाते हैं:

(1) पर्सनालिटी इनफार्मेशन गैदरिंग (प्रश्नावली) (Personality Information Gathering (Questionnaire)): डेटासेट का उपयोग प्रश्नावली तैयार

करने के लिए किया जाता है जो उपयोगकर्ता से टेक्स्ट प्राप्त करता है। यह टेक्स्ट JSON फाइल में संग्रहीत है और यह डेटा स्ट्रक्चर यानी डेटाफ्रेम (DataFrame) का उपयोग करके JSON फाइल में संग्रहीत है जो पांडा (Pandas) लाइब्रेरी के अंतर्गत मौजूद है।

(2) डेटा प्री–प्रोसेसिंग (मिन–मैक्स स्केलर) (Data Pre-Processing (Min-Max Scalar)): मिन–मैक्स स्केलर का उपयोग डेटा प्रीप्रोसेसिंग के लिए किया जाता है। इसका उपयोग मूलतः सामान्यीकरण (Normalization) के लिए किया जाता है। इस प्रक्रिया में डेटाफ्रेम से सूची के रूप में डेटा प्राप्त होता है और मिन–मैक्स एल्गोरिथ्म इसे संसाधित करता है और तदनुसार परिणाम प्रदान करता है। मिन–मैक्स स्केलिंग का आउटपुट नया डेटासेट है जहां प्रत्येक सुविधा का मान 0 से 1 के बीच स्केल किया जाता है और सूची में संग्रहीत किया जाता है [1]।

(3) डेटा संवर्धन (क्लस्टर ऑप्टिमाइजेशन) (Data Enrichment (Cluster Optimization)): क्लस्टर ऑप्टिमाइजेशन तकनीक का उपयोग डेटा संवर्धन उद्देश्य के लिए किया जाता है। सिल्हूट स्कोर (Silhouette score) का उपयोग क्लस्टर ऑप्टिमाइजेशन के लिए किया जाता है। क्लस्टरों की इष्टतम संख्या ज्ञात करने के लिए, क्लस्टर ऑप्टिमाइजेशन का उपयोग किया जाता है। क्लस्टरों की संख्या (k) चुनने के लिए, k क्लस्टर ऑप्टिमाइजेशन का आवश्यक पहलू है। इस चरण के लिए इनपुट पिछली स्थिति से सूची (list) है और आउटपुट k का एकल मान है।

(4) पैटर्न पहचान (के–मीन्स क्लस्टरिंग) (Pattern Recognition (K-Means Clustering)): क्लस्टरिंग एल्गोरिथ्म के–मीन्स का उपयोग अक्सर पैटर्न पहचान के लिए किया जाता है। डेटा बिंदुओं का डेटा सेट और क्लस्टर की संख्या यानी k इस चरण के लिए इनपुट हैं। के–मीन्स एल्गोरिथ्म इसे प्रोसेस करता है और प्रत्येक डेटा बिंदु के क्लस्टर के लिए आउटपुट यानी लेबल उत्पन्न करता है। यह चरण के–मीन्स क्लस्टरिंग का उपयोग करके व्यक्ति के पर्सनालिटी की भविष्यवाणी करता है।

(5) पर्सनालिटी लक्षण की पहचान (क्लस्टर परिवर्तन तकनीक) (Personality Trait Identification (Cluster Varying Technique)): पिछला चरण निश्चित क्लस्टर आकार के लिए आउटपुट उत्पन्न करता है। यह चरण, क्लस्टर आकार को 2 से 14 तक बदलता है। प्रत्येक क्लस्टर आकार व्यक्तिगत आउटपुट उत्पन्न करता है और व्यक्ति के पर्सनालिटी की भविष्यवाणी करता है।

(6) पर्सनालिटी प्रेडिक्शन (आउटपुट) (Personality Prediction (Output)): यह चरण व्यक्ति के पर्सनालिटी की भविष्यवाणी करता है। आउटपुट उत्पन्न करने के लिए वे पिछले चरण में उत्पन्न अधिकतम पर्सनालिटी वैल्यू पर विचार करते हैं। अधिकतम पर्सनालिटी वैल्यू की गणना करके, पर्सनालिटी की भविष्यवाणी की जाती है।

(7) डेटा एकत्रीकरण (Kaggle डेटासेट) (Data Gathering (Kaggle Dataset)): कागल (kaggle) ओपन डेटासेट का उपयोग बायोडाटा वर्गीकरण उद्देश्य के लिए किया है। टेक्स्ट को Comma-Seperated Values (CSV) फाइल के रूप में संग्रहीत किया है। पांडा (Pandas) लाइब्रेरी का उपयोग CSV फाइल को पढ़ने के लिए किया है और इसकी सामग्री डेटाफ्रेम में संग्रहीत की जाएगी।

(8) प्रशिक्षण डेटासेट पर टेक्स्ट वेक्टराइजेशन (Tf-Idf वेक्टराइजेशन) (Text Vectorization on Training Dataset (Tf-Idf Vecrorization)): यह डेटाफ्रेम को इनपुट के रूप में लेता है। इस डेटाफ्रेम में वह टेक्स्ट है जिसे संख्यात्मक डेटा में बदलने की आवश्यकता है। यह scikit-learn लाइब्रेरी से Tf-Idf Vectorizer क्लास

का उपयोग करके किया जा सकता है। fit() मेथड निष्पादित होने के बाद Tfidf Vectorizer ऑब्जेक्ट को प्रशिक्षित किया जाता है।

(9) नौकरी शीर्षक लेबल एनकोडर (scikit-learn लाइब्रेरी) (Job Title Encoder (Scikit-learn library)): यह चरण सभी श्रेणीगत वैल्यू को संख्यात्मक वैल्यू में परिवर्तित करता है। पायथन (python) की scikit-learn लाइब्रेरी का उपयोग लेबल एन्कोडिंग उद्देश्य के लिए किया जाता है। Label Encoder क्लास की transform मेथड इनपुट के रूप में एक श्रेणीगत विशेषता लेती है और संख्यात्मक वैल्यू लौटाती है। इस चरण का आउटपुट प्रत्येक श्रेणीगत विशेषता के संख्यात्मक वैल्यू की सूची है।

(10) pdf प्रारूप में इनपुट बायोडाटा (PyPDF2 पायथन लाइब्रेरी) (Input Resume in PDF Format (PyPDF2 library)): इस चरण में, उपयोगकर्ता से pdf के रूप में बायोडाटा लें। PyPDF2 पायथन लाइब्रेरी का उपयोग pdf रिज्यूमे से कंटेंट निकालने के लिए किया जाता है और निकाले गए डेटा को अगले चरण में भेजा जाता है [6]।

(11) डेटा प्री–प्रोसेसिंग (टेक्स्ट एक्सट्रेक्शन तकनीक) (Data Pre-Processing (Text Extraction Technique)): ये चरण इनपुट के रूप बायोडाटा लेता हैं और विभिन्न टेक्स्ट एक्सट्रेक्शन तकनीकों का उपयोग करके प्री–प्रोसेसिंग किया जाता है। विराम चिह्न हटाने के लिए पायथन की 're' लाइब्रेरी का उपयोग किया जाता है। स्टॉपवर्ड्स को हटाने के लिए पायथन की NLTK लाइब्रेरी का उपयोग किया जाता है। प्री–प्रोसेसिंग तकनीक लागू करने के बाद बायोडाटा क्लीन होता है और आगे के चरण के लिए भेज दिया जाता है [1]।

(12) परीक्षण डेटासेट पर टेक्स्ट वेक्टराइजेशन (Tf-Idf वेक्टराइजेशन) (Text Vectorization on Testing Dataset (Tf-Idf Vectorization)): इस चरण में, Tf-IdfVectorizer का प्रशिक्षित ऑब्जे. क्ट transform() मेथड का उपयोग करके साफ किए गए रेज़्यूमे पर ट्रांसफॉर्म होता है। इस चरण का आउटपुट फीचर मैट्रिक्स है जिसका उपयोग आगे वर्गीकरण के लिए किया जाता है [5]।

(13) नौकरी शीर्षक भविष्यवाणी (RF क्लासिफायर) (Job Title Prediction (RF Classifier)): इस चरण में, पिकल लाइब्रेरी का उपयोग करके RF क्लासिफायर मॉडल को लोड करें। आगे के चरण में, predict() मेथड का उपयोग करके प्रशिक्षित मॉडल को क्लास लेबल का प्रेडिक्शन करने के लिए फीचर मैट्रिक्स पर लागू किया जाता है। रिवर्स एन्कोडिंग inverse–transform() मेथड का उपयोग करके भविष्यवाणी की जाती है यानी संख्यात्मक मान श्रेणीबद्ध मान में परिवर्तित होता है [3]।

प्रस्तावित प्रणाली के लिए निम्नलिखित एल्गोरिथ्म का उपयोग किया जाता है।

एल्गोरिथ्म 1 : पर्सनालिटी का पता लगाना

इनपुट : क्लस्टर की संख्या 'k', डेटासेट की संख्या 'x'

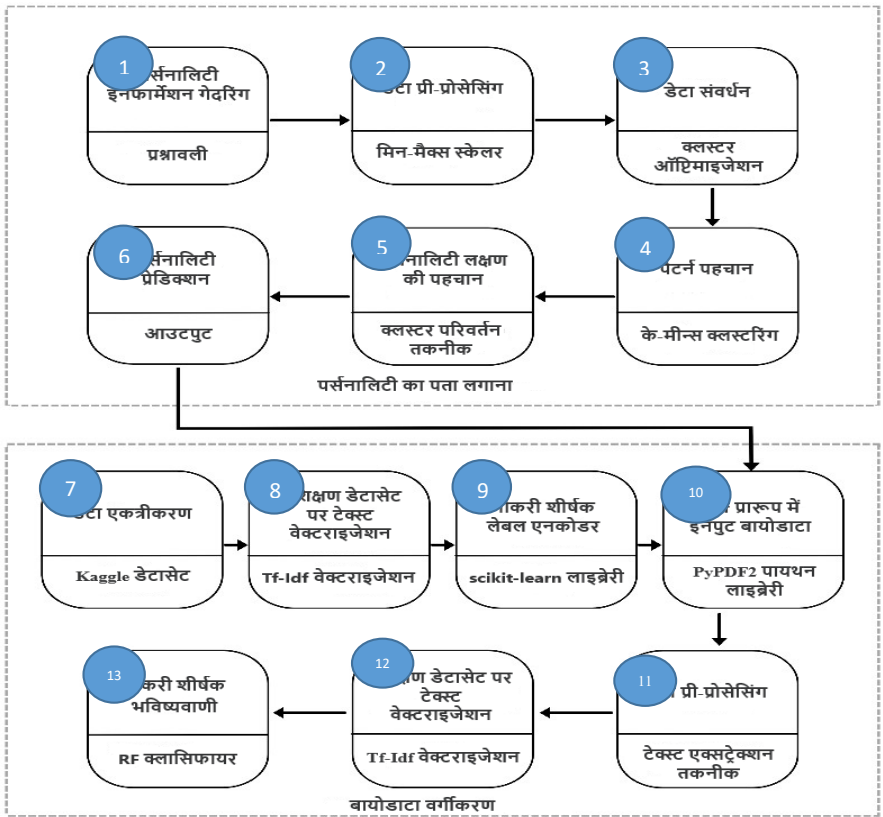
आउटपुट : प्रत्येक क्लस्टर को लेबल असाइन किया

चरण 1 : उपयोगकर्ता से Q और A प्रारूप में इनपुट लें और डेटाफ्रेम का उपयोग करके JSON फाइल में संग्रहीत करें।

चरण 2 : 0 से 1 के बीच रूपांतरित डेटा पर मिन–मैक्स स्केलिंग लागू करें।

चरण 3 : इनिशियल सेंट्रोइड के रूप में डेटासेट से k डेटा बिंदुओं की वैल्यू चुनें।

चरण 4 : प्रत्येक डेटा बिंदु को निकटतम माध्य (mean) वाले क्लस्टर को सौंपा जाना चाहिए।



चित्र 1: प्रस्तावित सिस्टम फ्रेमवर्क

चरण 5 : क्लस्टर mean को अपडेट करें।

चरण 6 : चरण 4 और 5 को तब तक दुहराएँ जब तक कि क्लस्टर अलोकेशन में कोई बदलाव न हो जाए।

चरण 7 : अंतिम सेंट्रोइड और क्लस्टर असाइनमेंट की वैल्यू लौटाएँ यानी प्रत्येक क्लस्टर को लेबल सौंपा गया।

एल्गोरिथ्म 2: बायोडाटा वर्गीकरण

इनपुट : फीचर्स का सेट

आउटपुट : प्रत्येक डेटा पॉइंट का लेबल

चरण 1 : उपयोगकर्ता से pdf रूप में इनपुट के रूप में बायोडाटा लें।

चरण 2 : स्टॉपवर्ड और विराम चिह्न हटाने के लिए NLTK लाइब्रेरी और रेगुलर एक्सप्रेशन का उपयोग करें।

चरण 3 : असंरचित रेज़्यूमे को क्लीन रेज़्यूमे में बदलने के लिए Tfidf Vectorizer अप्लाई करें।

चरण 4 : ट्री की संख्या स्पेसिफाय करके एक यादृच्छिक रैंडम फॉरेस्ट बनाएं।

चरण 5 : प्रशिक्षण डेटा पर मॉडल फिट करें।

चरण 6 : परीक्षण डेटा पर एक प्रेडिक्शन करें।

परिणाम और चर्चा

प्रस्तावित प्रणाली फ्रेमवर्क के तहत, व्यक्तित्व पहचान से संबंधित विभिन्न प्रश्नावली की पहचान की जाती है जिसके लिए प्रश्नावली को अंतिम रूप देने के लिए विभिन्न लेखों और पत्रिकाओं का उपयोग किया जाता है [10]। वेब प्रौद्योगिकी का उपयोग करते हुए यह प्रश्नावली उपयोगकर्ता को उनके व्यक्तिगत लॉगिन के बाद प्रदान की जाती है। एक बार उपयोगकर्ता लॉग इन हो जाने पर, सभी श्रेणियों पर आधारित प्रश्नावली प्रदर्शित की जाएंगी। निम्नलिखित व्यक्तित्व पहचान वाली कुल पाँच श्रेणियाँ हैं: i) बहिर्मुखता (Extroversion), ii) मनोविक्षुब्धता (Neuroticism), iii) सहमतता (Agreeableness), iv) कर्तव्यनिष्ठा (Conscientiousness), अ) खुलापन (Openness)। प्रत्येक श्रेणी को व्यक्तिगत विशेषताओं के अनुसार पहचानने की आवश्यकता है। प्रत्येक श्रेणी के अंतर्गत 10 महत्वपूर्ण प्रश्नावली की पहचान की जाती है और उपयोगकर्ता को समान पैमाने के आधार पर प्रदान की जाती है। लाइकर्ट स्केल 1 से 6 तक है। लाइकर्ट स्केल का वैल्यू इस प्रकार दिया गया है: 1 अज्ञात, 2 आंशिक रूप से असहमत, 3 असहमत, 4 तटस्थ, 5 आंशिक रूप से सहमत और 6 सहमत। निम्नलिखित टेबल 1 प्रश्नावली प्रदान करता है।

टेबल 1 : व्यक्तित्व पहचान प्रश्नावली–आधारित प्राप्त नमूना इनपुट

वर्गी – करण	क्रम संख्या	प्रश्नावली	स्केल (1–>अज्ञात, 2–>आंशिक रूप से असहमत, 3–>असहमत, 4–>तटस्थ, 5–>आंशिक रूप से सहमत, 6–>सहमत)					
			1	2	3	4	5	6
बहिर्मुखता (Extroversion)	1	मैं पार्टी की जान हूँ।				■		
	2	मैं ज्यादा बात नहीं करता हूँ।			■			
	3	मैं लोगों के बीच सहज महसूस करता हूँ।				■		
	4	मैं पीछे रहता हूँ।				■		
	5	मैं बातचीत शुरू करता हूँ।				■		
	6	मेरे पास कहने को बहुत कम है।					■	
	7	मैं पार्टियों में कई अलग–अलग लोगों से बात करता हूँ।				■		
	8	मुझे अपनी ओर ध्यान आकर्षित करना पसंद नहीं है।				■		
	9	मुझे आकर्षण का केंद्र बनने से कोई परेशानी नहीं है।						■
	10	मैं अजनबियों के आसपास शांत रहता हूँ।						■

मनोविक्षुब्धता (Neuroticism)	1	मैं आसानी से तनावग्रस्त हो जाता हूँ।			■			
	2	मैं अधिकांश समय तनावमुक्त रहता हूँ।				■		
	3	मुझे चीजों की चिंता है।			■			
	4	मुझे शायद ही कभी दुःख महसूस होता है।			■			
	5	मैं आसानी से डिस्टर्ब हो जाता हूँ।		■				
	6	मैं आसानी से परेशान हो जाता हूँ।				■		
	7	मैं अपना मूड बहुत बदलता हूँ।		■				
	8	मेरा मूड बार–बार बदलता रहता है।		■				
	9	मैं आसानी से चिढ़ जाता हूँ।		■				
	10	मुझे अक्सर दुःख होता है।				■		
सहमतता (Agreeableness)	1	मुझे दूसरों के प्रति थोड़ी चिंता महसूस होती है।						■
	2	मुझे लोगों में दिलचस्पी है।						■
	3	मैं लोगों का अपमान करता हूँ।		■				
	4	मुझे दूसरों की भावनाओं से सहानुभूति है।						■
	5	मुझे दूसरे लोगों की समस्याओं में कोई दिलचस्पी नहीं है।			■			
	6	मेरा हृदय कोमल है।						■
	7	मुझे वास्तव में दूसरों में कोई दिलचस्पी नहीं है।				■		
	8	मैं दूसरों के लिए समय निकालता हूँ।						■
	9	मैं दूसरों की भावनाओं को महसूस करता हूँ।						■
	10	मैं लोगों को सहज महसूस कराता हूँ।				■		
कर्तव्यनिष्ठा (Conscientiousness)	1	मैं हमेशा तैयार रहता हूँ।						■
	2	मैं अपना सामान इधर–उधर छोड़ देता हूँ।			■			
	3	मैं विवरण पर ध्यान देता हूँ।						■
	4	मैं गड़बड़ी करता हूँ।		■				
	5	मैं काम जल्द ही कर लेता हूँ।				■		
	6	मैं अक्सर चीजों को उनके उचित स्थान पर वापस रखना भूल जाता हूँ।					■	
	7	मुझे अपने वरिष्ठ द्वारा दिया गया आदेश पसंद है।		■				
	8	मैं अपने कर्तव्यों से भागता हूँ।			■			
	9	मैं शेड्यूल का पालन करता हूँ।						■
	10	मैं अपने काम में सटीक हूँ।					■	

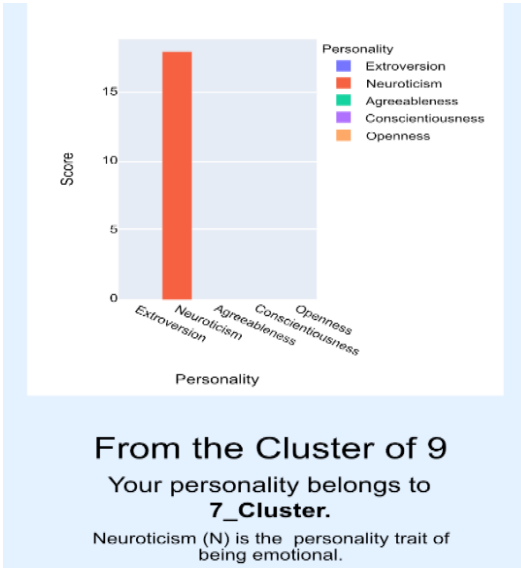
खुलापन (Openness)	1	मेरे पास एक समृद्ध शब्दावली है।					■	
	2	मुझे अमूर्त विचारों को समझने में कठिनाई होती है।				■		
	3	मेरे पास एक ज्वलंत कल्पना है।				■		
	4	मुझे अमूर्त विचारों में कोई दिलचस्पी नहीं है।		■				
	5	मेरे पास बेहतरीन विचार है।						■
	6	मेरी कल्पनाशक्ति अच्छी नहीं है।		■				
	7	मैं चीजों को जल्दी समझ लेता हूँ।						■
	8	मैं कठिन शब्दों का प्रयोग करता हूँ।		■				
	9	मैं चीजों पर सोचने में समय बिताता हूँ।			■			
	10	मैं विचारों से भरा हूँ।						■

टेबल 2: क्लस्टर आकार 2 से 14 तक मॉड्यूल 1 का आउटपुट

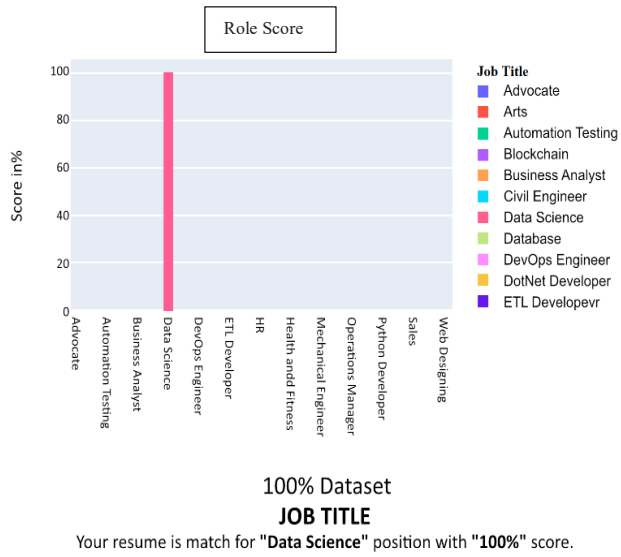
क्लस्टर का आकार	व्यक्तित्व क्लस्टर	व्यक्तित्व प्रकार
2	क्लस्टर 1	अनुभव के लिए खुलापन (Openness to Experience)
3	क्लस्टर 2	अनुभव के लिए खुलापन (Openness to Experience)
4	क्लस्टर 1	अनुभव के लिए खुलापन (Openness to Experience)
5	क्लस्टर 3	अनुभव के लिए खुलापन (Openness to Experience)
6	क्लस्टर 3	अनुभव के लिए खुलापन (Openness to Experience)
7	क्लस्टर 3	अनुभव के लिए खुलापन (Openness to Experience)
8	क्लस्टर 4	अनुभव के लिए खुलापन (Openness to Experience)
9	क्लस्टर 4	अनुभव के लिए खुलापन (Openness to Experience)
10	क्लस्टर 2	अनुभव के लिए खुलापन (Openness to Experience)
11	क्लस्टर 1	अनुभव के लिए खुलापन (Openness to Experience)
12	क्लस्टर 10	अनुभव के लिए खुलापन (Openness to Experience)
13	क्लस्टर 4	अनुभव के लिए खुलापन (Openness to Experience)
14	क्लस्टर 3	अनुभव के लिए खुलापन (Openness to Experience)

टेबल 3: मॉड्यूल 2 का आउटपुट (डेटा सेट का आकार 100% से 40% तक)

डेटासेट का आकार	नौकरी शीर्षक	स्कोर
100%	डेटा वैज्ञानिक	50.0
60%	डेटा वैज्ञानिक	28.57
40%	डेटा वैज्ञानिक	14.29



चित्र 2: मॉड्यूल 1 का आउटपुट (क्लस्टर आकार 9)



चित्र 3: 100% डेटासेट पर विचार करते हुए मॉड्यूल 2 का आउटपुट

यह पाया गया है कि यदि क्लस्टर का आकार बदला, तब अधिकतम पहचाने गए व्यक्तित्व पर विचार किया जाता है। अधिकतम व्यक्तित्व मूल्य अगले स्तर के लिए लागू होता है। दूसरे मॉड्यूल में रेज़्यूमे वर्गीकरण में न केवल पीडीएफ प्रारूप में विभिन्न नौकरी चाहने वालों के रेज़्यूमे से इनपुट लिया जा रहा है, बल्कि स्वयं के व्यक्तित्व की पहचान के लिए मॉड्यूल 1 के तहत प्रदर्शन परिणाम पर भी विचार किया जाता है और संसाधित किया जाता है। अगली स्थिति पीडीएफ प्रारूप में सारांश इनपुट करना है। pdf फाइल को पढ़ने के लिए PyPDF2 लाइब्रेरी का उपयोग किया जाता है। अगला चरण प्री–प्रोसेसिंग है जो NLTK लाइब्रेरी और रेगुलर एक्सप्रेशन का उपयोग करके किया जाता है। एक बार बायोडाटा साफ हो जाने पर, Tfidf Vectorizer के प्रशिक्षित ऑब्जेक्ट का उपयोग करके इसे ट्रांसफॉर्म किया जाता है। यह फीचर मैट्रिक्स उत्पन्न करता है जो आगे के वर्गीकरण के लिए आवश्यक है। अगले चरण में, पिकल लाइब्रेरी का उपयोग होता है, आरएफ क्लासिफायर मॉडल को लोड करके और predict() विधि का उपयोग करके क्लास लेबल की भविष्यवाणी होती है। अंत में, संख्यात्मक मान को श्रेणीबद्ध मान में बदलने के लिए व्युत्क्रम–परिवर्तन() विधि का उपयोग होता है। प्रस्तावित प्रणाली चित्र 3 प्रदान करती है जो नौकरी शीर्षक भविष्यवाणी के आधार पर सारांश वर्गीकरण का ग्राफिकल आउटपुट है जो 100% डेटासेट पर विचार करता है। टेबल 3 समग्र टेबल को दर्शाती है जब डेटासेट का आकार 100% से 40% तक भिन्न होता है। चित्र 3 विभिन्न नौकरी शीर्षकों के लिए बायोडाटा का अलग–अलग स्कोर प्रदर्शित करता है। 100% से 40% के बीच डेटासेट का आकार बदलने के बाद उच्चतम स्कोर वाले अधिकतम नौकरी शीर्षक नाम पर अंतिम रूप से विचार किया जाएगा। उसके बाद आगे की सटीकता पहचान के उद्देश्य से व्यक्तित्व का पता लगाने और नौकरी के शीर्षक की पहचान के लिए 5 साल से अधिक का अनुभव रखने वाले 10 डेटा वैज्ञानिक व्यक्तियों का चयन किया है। निम्नलिखित टेबल 4 सफलता अनुपात के साथ 10 व्यक्तियों का आउटपुट देता है।

टेबल 4: 10 व्यक्तियों की नौकरी के शीर्षक की भविष्यवाणी

व्यक्ति	डेटासेट का आकार	नौकरी शीर्षक	टिप्पणी
व्यक्ति 1	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	डेटाबेस	
व्यक्ति 2	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	डेटा वैज्ञानिक	
व्यक्ति 3	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	यांत्रिक इंजीनियर	
व्यक्ति 4	100%	पायथन डेवलपर	असफल
	60%	डेटा वैज्ञानिक	
	40%	पायथन डेवलपर	
व्यक्ति 5	100%	डेटा वैज्ञानिक	सफल
	60%	डेवऑप्स इंजीनियर	
	40%	डेटा वैज्ञानिक	
व्यक्ति 6	100%	डेटाबेस	असफल
	60%	डेटाबेस	
	40%	डेटा वैज्ञानिक	
व्यक्ति 7	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	डेटा वैज्ञानिक	
व्यक्ति 8	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	संचालन प्रबंधक	
व्यक्ति 9	100%	व्यापार विश्लेषक	असफल
	60%	डेटा वैज्ञानिक	
	40%	व्यापार विश्लेषक	
व्यक्ति 10	100%	डेटा वैज्ञानिक	सफल
	60%	डेटा वैज्ञानिक	
	40%	पायथन डेवलपर	

इसलिए 10 व्यक्तियों में से 7 व्यक्तियों का नौकरी शीर्षक सही है और 3 व्यक्तियों का गलत है। अतः इस प्रणाली की सटीकता 70% है। सटीकता बढ़ाने के लिए एन्सेम्बल लर्निंग तकनीक या रैखिक एसवीसी का उपयोग किया जा सकता है।

5. निष्कर्ष और भविष्य की संभावनाएँ

प्रस्तावित प्रणाली नियुक्ति प्रक्रिया को तेज बनाती है और इसमें कम समय और प्रयास की आवश्यकता होती है। व्यक्ति के व्यक्तित्व के अनुसार विशिष्ट नौकरी की स्थिति के लिए बायोडाटा को वर्गीकृत करने के लिए, के–मीन्स और आरएफ का प्रभाव विश्लेषण उपयोगी दृष्टिकोण हैं। यह मॉडल दो मॉड्यूल पर आधारित है:

i) व्यक्तित्व का पता लगाना जो के–मीन्स क्लस्टरिंग एल्गोरिथ्म का उपयोग करके व्यक्ति के व्यक्तित्व का पता लगाता है।

ii) बायोडाटा वर्गीकरण जिसका उपयोग आरएफ वर्गीकरण मॉडल का उपयोग करके नौकरी के प्रकार की पहचान करने के लिए किया जाता है। परिणाम विश्लेषण के अनुसार, मॉडल उचित तरीके से कार्य करता प्रतीत होता है और 70% सटीकता के साथ व्यक्ति के नौकरी शीर्षक की भविष्यवाणी करता है। भविष्य में, एन्सेम्बल लर्निंग तकनीक या लीनियर एसवीसी का उपयोग करके प्रस्तावित मॉडल की सटीकता में सुधार किया जा सकता है। सभी प्रकार की फाइल का सारांश लेकर इस मॉडल को बेहतर बनाया जा सकता है।

शोध पत्र में प्रयुक्त अंग्रेजी शब्दों की समानार्थक हिंदी शब्दावली

Alphabetically sorted terminology in English	वर्णमाला अनुक्रमित हिंदी शब्दावली
Agreeableness	सहमतता
Conscientiousness	कर्तव्यनिष्ठा

Data enrichment	डेटा संवर्धन
Data gathering	डेटा एकत्रीकरण
Extroversion	बहिर्मुखता
Natural Language Processing (NLP)	प्राकृतिक भाषा प्रसंस्करण
Neuroticism	मनोविक्षुब्धता
Openness	खुलापन
Optical Character Recognition (OCR)	प्रकाशिक संप्रतीक अभिज्ञान
Pattern Recognition	पैटर्न पहचान

संदर्भः

[1] P. Swami and V. Pratap, "Resume Classifier and Summarizer," in 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022.

[2] D. Reddy , S. Regella and S. Seelam , "Recruitment Prediction using Machine Learning," in 5th International Conference on Computing, Communication and Security (ICCCS), 2020.

[3] S. Amin, N. Jayakar, S. Sunny, P. Babu, M. Kiruthika and A. Gurjar, "Web Application for Screening Resume," in International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India., 2019.

[4] I. H. Sarkar, "Machine Learning: Algorithms, Real-World Applications and Research Directions," SN Computer Science, vol. 2, no. 3, p. 160, 2021.

[5] S. Ramraj, V. Sivakumar and R. . G. Kaushik, "Real-Time Resume Classification System Using LinkedIn Profile Descriptions," in 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), Keonjhar, India, 2020.

[6] B. Gunaseelan, S. Mandal and V. Rajagopalan, "Automatic Extraction of Segments from Resumes using Machine Learning," in 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, 2020.

[7] P. Roy , S. Chowdhary and R. Bhatia, "A Machine Learning approach for automation of Resume Recommendation system," Procedia Computer Science, vol. 167, pp. 2318-2327, 2020/01/01.

[8] C. Maddumage, D. Senevirathne, . I. Gayashan, . T. Shehan and S. Sumathipala, "Intelligent Recruitment System," in 2019 5th International Conference for Convergence in Technology (ICCT) , Pune, India, 2019.

[9] R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," in 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019.

[10] L. R. Goldberg, "The development of markers for the Big-Five factor structure.," Psychological Assessment, vol. 4, no. 1, pp. 26-42, Mar 1992.