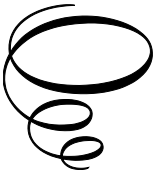
Health Informatics Using Artificial Intelligence

Health Informatics Using Artificial Intelligence

Edited by

Prabhpreet Kaur and Amandeep Kaur

Cambridge Scholars Publishing



Health Informatics Using Artificial Intelligence

Edited by Prabhpreet Kaur and Amandeep Kaur

This book first published 2026

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-6944-3
ISBN (Ebook): 978-1-0364-6945-0

TABLE OF CONTENTS

EDITORS AND CONTRIBUTORS

Editors

Dr. Prabhpreet Kaur is an Associate Professor in Guru Nanak Dev University (GNDU), Amritsar, Punjab. She did her Ph.D. in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. She received her M.Tech. (Computer Science and Engineering) from Guru Nanak Dev Engineering College (Ludhiana) affiliated to Punjab Technical University, Jalandhar. She has 15 Years of extensive Teaching Experience at Graduate level. She has attended and conducted various Seminars, Conferences, and Workshops and FDP etc. She has published various research papers in reputed International Journals (SCI/SCIE/ESCI/Scopus etc. indexed) and Conference Proceedings (IEEE and Springer etc.). Her major interest includes Medical Imaging, Machine Learning, Deep Learning using Python and Matlab, and Artificial Intelligence. Her research has been cited 1048 times, and she holds an H-index of 16.

Dr. Amandeep Kaur received her B. Tech degree in Computer Science and Engineering from Punjab Technical University Jalandhar, Punjab, India, in 2008, her Master's degree in Computer Science and Engineering from Guru Nanak Dev University, Punjab, India in 2011 and her Ph.D. from Guru Nanak Dev University Amritsar, Punjab, India in 2023.She has 14 years of teaching and research experience. She is an Assistant Professor at the Department of Computer Engineering and Technology (CET), Guru Nanak Dev University, Amritsar, Punjab, India. To her credit, she has more than 25 research papers in international journals and proceedings of international conferences. Her areas of interest include learning analytics, artificial intelligence and AI-based smart healthcare and medical disease diagnostic applications.

Chapter 1: Machine Learning Models for Predicting the Success of Health Care Apps on Google Play Store

Author Contribution:

Anureet Kaur contributed to the development and implementation part. She was involved in performance evaluation of the proposed framework.

Dr. Anureet Kaur is currently serving as an Assistant Professor in the Department of Computer Science and Applications at Khalsa College, Amritsar, since 2016. She brings with her a rich experience of over 16 years in teaching and research, dedicated to nurturing future technologists and contributing to the advancement of computer science education.

She has authored more than 30 research papers in reputed national and international journals and conferences, and has contributed two book chapters in the areas of emerging technologies. Her research interests span across Software Engineering, Mobile Application Development, and Machine Learning, with a particular focus on practical implementations and innovation-driven solutions.

Chapter 2: Deep Transfer Learning for Brain Tumor Classification: Boosting Accuracy with Advanced Classifiers

Author Contribution:

Priyanka Mahajan contributed to the development and implementation of deep learning models for medical image classification. She was involved in data preprocessing, feature extraction, and performance evaluation of the proposed framework.

Prabhpreet Kaur contributed to the conceptualization and experimental validation of the proposed methodology. She assisted in dataset preparation, model optimization, and analysis of classification results.

Suman contributed to the proof-reading part of paper. She was involved in drafting of the paper.

Priyanka Mahajan, M.Tech. (CSE), B.Tech. (CSE)

Priyanka Mahajan is an Assistant Professor in Lovely Professional University (LPU), Phagwara, Punjab. She is pursuing her Ph.D. in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. She received her M.Tech. (Computer Science and Engineering) from GNDU (Guru Nanak Dev University). She has approximately 10 Years of Teaching Experience at any University. Her major interest includes Medical Imaging, Machine Learning, Deep Learning using Python.

Suman Bala, M.Tech. (CSE), B.E. (IT)

Suman Bala is an Assistant Professor in Guru Nanak Dev University, Amritsar, Punjab. She is pursuing her Ph.D. in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. She received her M.Tech. (Computer Science and Engineering) from GNDU (Guru Nanak Dev University). Her major interest includes latest techniques of Machine Learning and Deep Learning using Python.

Chapter 3: Cervical Cell Segmentation and Classification by Leveraging SE-UNet and DenseCapsNet Models

Author Contribution:

Priyanka Mahajan contributed to the development and implementation of deep learning models for medical image classification. She was involved in data preprocessing, feature extraction, and performance evaluation of the proposed framework.

Prabhpreet Kaur contributed to the conceptualization and experimental validation of the proposed methodology. She assisted in dataset preparation, model optimization, and analysis of classification results.

Priyanka Mahajan, M.Tech. (CSE), B.Tech. (CSE)

Priyanka Mahajan is an Assistant Professor in Lovely Professional University (LPU), Phagwara, Punjab. She is pursuing her Ph.D. in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. She received her M.Tech. (Computer Science and Engineering)

from GNDU (Guru Nanak Dev University). She has approximately 10 Years of Teaching Experience at any University. Her major interest includes Medical Imaging, Machine Learning, Deep Learning using Python.

Chapter 4: GANs for Synthetic Medical Data: Enhancing Data Availability, Privacy, and Model Performance in Healthcare AI

Author Contribution:

Study conception and design: Radhika Kumari

Data collection: Kiranbir Kaur

Analysis and interpretation of results: Kuldeep Kumar

Dr. Radhika Kumari is currently pursuing her Ph.D. in Computer Engineering & Technology, at Guru Nanak Dev University, Amritsar, India. Her primary research interests lie in the fields of Machine Learning and the Internet of Things (IoT), with a focus on developing intelligent, data-driven solutions for real-world problems. She has authored and co- authored numerous research papers in reputed national and international journals and conferences. Her academic work reflects a strong commitment to innovation, interdisciplinary collaboration, and technological advancement. Radhika continues to contribute to the scientific community through her research and active participation in scholarly activities.

Er. Kiranbir Kaur is an Assistant Professor at Department of Computer Engineering & Technology, Guru Nanak Dev University, Amritsar, India. She has completed M.Tech (CSE) and holds PhD from Guru Nanak Dev University, Amritsar. She has cleared UGC-NET JRF exam. Her area of interest includes Cloud Computing, databases and wireless sensor networks. Her work on these topics has been published in a range of international journals including ACM Computing Surveys, Journal of Supercomputing, Applied Sciences and Sensors. She has more than 14 years of experience in teaching.

Er. Kuldeep Kumar is a passionate researcher pursuing a Ph.D. in Computer Science and Engineering at Chandigarh University, Punjab. He has published several research papers in reputed national and international

conferences, contributing to emerging technologies. His key interests include blockchain security, AI, machine learning, and IoT-based smart systems. Known for his technical depth and originality, his work focuses on practical, real-world applications. In addition to his academic achievements, he holds six granted patents, reflecting his innovative mindset and commitment to solving real-world problems.

Chapter 5: Advancements in Deepfake Detection using Deep Learning: An Overview

Author Contribution:

Study conception and design: Vishaldeep Kaur, Pawandeep Kaur

Data collection: Kiranbir Kaur, Prabhpreet Kaur

Analysis and interpretation of results: Vishaldeep Kaur, Pawandeep Kaur, Kiranbir Kaur, Prabhpreet Kaur

Drafting of manuscript: Vishaldeep Kaur, Pawandeep Kaur

All authors reviewed the results and approved the final version of the manuscript.

Vishaldeep Kaur, M.Tech(CSE), B.Tech(CSE)

Vishaldeep Kaur has completed M.Tech in Computer Science and Engineering from the Department of Computer Engineering and Technology at Guru Nanak Dev University(GNDU), Amritsar, Punjab. She received her B.Tech degree in Computer Science and Engineering from the same institution. Her academic interests include Deep Learning, Machine Learning and Multimedia Forensics. Her current work aims to contribute to the development of efficient and interpretable AI-based solutions in the domain of multimedia forensics. She has published a research article in Scopus indexed journal.

Dr. Kiranbir Kaur is an Assistant Professor at Department of Computer Engineering & Technology, Guru Nanak Dev University, Amritsar, India. She has completed M.Tech (CSE) and holds PhD from Guru Nanak Dev

University, Amritsar. She has cleared UGC-NET JRF exam. Her area of interest includes Cloud Computing, databases and wireless sensor networks. Her work on these topics has been published in a range of international journals including ACM Computing Surveys, Journal of Supercomputing, Applied Sciences and Sensors. She has more than 14 years of experience in teaching.

Pawandeep Kaur, M.Tech(CSE), B.Tech(ECE)

Pawandeep Kaur has completed her Master of Technology (M.Tech.) in Computer Science and Engineering from Guru Nanak Dev University, Punjab, India. Her academic interests include artificial intelligence, deep learning, Python and computer vision. She has been actively involved in research related to medical image analysis and is committed to contributing innovative solutions in the field of intelligent systems and healthcare technologies. She has published research article in Scopus indexed journal.

Chapter 6: Machine Learning and Deep learning in Breast Cancer Diagnosis: An In-Depth Survey

Author Contribution:

Pawandeep Kaur: Conceptualization, literature search, data curation, drafting of the original manuscript, and preparation of figures and tables.

Vishaldeep Kaur: Literature review, data analysis, manuscript writing, and critical revision of the draft.

Prabpreet Kaur: Supervision, methodology guidance, validation of the reviewed literature, and critical editing of the manuscript.

Kiranbir kaur: Supervision, project administration, conceptual guidance, and final review and approval of the manuscript.

All authors read and approved the final version of the manuscript.

Pawandeep Kaur, M.Tech(CSE), B.Tech(ECE)

Pawandeep Kaur has completed her Master of Technology (M.Tech.) in Computer Science and Engineering from Guru Nanak Dev University,

Punjab, India. Her academic interests include artificial intelligence, deep learning, Python and computer vision. She has been actively involved in research related to medical image analysis and is committed to contributing innovative solutions in the field of intelligent systems and healthcare technologies. She has published research article in Scopus indexed journal.

Vishaldeep Kaur, M.Tech(CSE), B.Tech(CSE)

Vishaldeep Kaur has completed M.Tech in Computer Science and Engineering from the Department of Computer Engineering and Technology at Guru Nanak Dev University (GNDU), Amritsar, Punjab. She received her B.Tech degree in Computer Science and Engineering from the same institution. Her academic interests include Deep Learning, Machine Learning and Multimedia Forensics. Her current work aims to contribute to the development of efficient and interpretable AI-based solutions in the domain of multimedia forensics. She has published a research article in Scopus indexed journal.

Kiranbir Kaur, Ph.D. (CSE), M.Tech(CSE), B.Tech(CSE)

Dr. Kiranbir Kaur is Assistant Professor in the Department of Computer Engineering and Technology at Guru Nanak Dev University (GNDU), Amritsar, Punjab. She received her Ph.D. and M.Tech. degrees in Computer Science and Engineering from Guru Nanak Dev University, Amritsar. With over 12 years of teaching and research experience, her academic interests include cloud computing, wireless sensor networks, big data, and energy-efficient algorithms. She has published various research papers in reputed International Journals (SCI/SCIE/ESCI/Scopus etc. indexed) and Conference Proceedings (IEEE and Springer etc.).

Chapter 7: A Bibliometric and Technological Review of Cervical Cancer Diagnostics

Author Contribution:

Sapna Arora: Conceptualization, literature search, data curation, drafting of the original manuscript, and preparation of figures and tables, Literature review, data analysis, manuscript writing, and critical revision of the draft.

Amandeep Kaur: Supervision, methodology guidance, validation of the reviewed literature, and critical editing of the manuscript. Supervision, project administration, conceptual guidance, and final review and approval of the manuscript.

All authors read and approved the final version of the manuscript.

Sapna Arora, M.Tech (CSE), B.Tech (CSE)

Sapna Arora has completed her B.Tech degree in Computer Science and Engineering from the Global group of institutes affiliated to Punjab Technical University, Master of Technology (M.Tech.) in Computer Science and Engineering from Guru Nanak Dev University, Punjab, India. Her academic interests include artificial intelligence, deep learning, machine learning, Java, Web development, and Python. She has been actively involved in research related to medical image analysis and is committed to contributing innovative solutions in the field of intelligent systems and healthcare technologies. She has published a research article in IEEE indexed journal.

Chapter 8: A Review and Critical Analysis of Deep Learning Techniques in Cancer Detection and Prediction

Author Contribution:

Dr Amardeep Gupta contributed to the conceptualization of the study, supervision of the research work, methodology design, and critical review of the manuscript. He also provided domain expertise in artificial intelligence and healthcare applications.

Dr Naina Handa contributed to the literature review, data collection and analysis, preparation of tables and comparative studies, drafting of the manuscript, and final editing of the chapter.

Dr. Amardeep Gupta is currently working as a faculty member at DAV College, Amritsar. His research interests include Artificial Intelligence, Machine Learning, Data Analytics, and Healthcare Informatics. He has published several research papers in national and international journals and conferences related to emerging technologies in computing.

Dr. Naina Handa is a faculty member at DAV College, Amritsar. Her research interests include Artificial Intelligence, Deep Learning, Data Science, and Medical Image Analysis. She has contributed to multiple research publications in the field of computer science and health technologies.

Chapter 9: AI-Based Analysis of COVID-19 Trends in India for the Year 2022

Author contribution:

Dr. Sarbjit Singh wrote the manuscript, performed computational work, and is the first author.

Dr. Jatinder Kaur conceived the idea, carried out computational work, and is the corresponding author.

Dr. Kulwinder Singh Parmar conducted the analysis.

Dr. Sarbjit Singh

Dr. Sarbjit Singh is a distinguished researcher and academician at Guru Nanak Dev University College, Narot Jaimal Singh, Pathankot, Punjab, India. His expertise lies in Wavelets, Time Series Modeling, Autoregressive Forecasting, Soft Computing Techniques, and Machine Learning. With 8 years of research experience and over 20 years of strong academic teaching experience, Dr. Sarbjit Singh has made significant contributions to his field, publishing 14 international papers in high-impact factor journals and 4 book chapters. His research focuses on developing hybrid modeling techniques using Wavelets and Data-Driven Models to forecast non- linear, chaotic, and complex systems. He has also authored a national book on Linear Algebra, further disseminating knowledge in his field. As a research supervisor, he guides a Ph.D. research scholar in the Department of Mathematics at Guru Nanak Dev University, Amritsar. With nearly 600 citations of his research publications, his work has garnered global recognition. With a passion for research and a commitment to excellence, he is a respected figure in the academic community, solidifying his position as a leading expert in the field.

Dr. Jatinder Kaur

Dr. Jatinder Kaur is a renowned researcher and academician at Guru Nanak Dev University College Verka Amritsar. Her area of interest and expertise is stochastic modeling and forecasting. With a strong academic background, Dr. Jatinder Kaur has made significant contributions to the field, publishing 10 international papers in high-impact factor journals. Her research expertise lies in developing and applying stochastic models to forecast complex systems. She has also authored chapters in several international edited books, further disseminating knowledge in the field. Her work has been recognized globally, and she continues to advance the field of stochastic modeling and forecasting. With a passion for research and a commitment to excellence, she is a respected figure in the academic community. Her dedication to exploring new frontiers in stochastic modeling has led to impactful publications and collaborations, solidifying her position as a leading expert in the field.

Dr. Kulwinder Singh Parmar

Dr. Kulwinder Singh Parmar is a distinguished academician and researcher in mathematics. With a Ph.D. from Guru Gobind Singh Indraprastha University, he has over 14 years of teaching experience and has held positions at prominent institutions. Dr. Parmar's research focuses on stochastic modeling, water quality, and pollution modeling, with 60 publications in esteemed International Journals with high Impact factors. He has presented research papers at international conferences and collaborated with institutions like CSIR-National Physical Laboratory. Dr. Parmar has received several awards, including the Graduate Scholar Award at the University of British Columbia, Canada, and the State Award for Social Work from Bharat Scouts & Guides. He has also been granted fellowships, including a UGC Fellowship and a DST Travel Grant to Ohio State University, USA. Dr. Parma's international exposure includes visits to universities in the USA, Spain, and Canada, solidifying his expertise and reputation in the field of mathematics.

CHAPTER 1

MACHINE LEARNING MODELS FOR PREDICTING THE SUCCESS OF HEALTH CARE APPS ON GOOGLE PLAY STORE

ANUREET KAUR, PRABHPREET KAUR, AMANDEEP KAUR

Abstract

Background: At present, mobile apps play an imperative role in the lives of many people. Tens of thousands of apps are being put on the Play Store for users to install. The utilization of mobile technology in medicine and healthcare has become a decisive aspect. Mobile health (mHealth) is now referred to as the field related to health-related apps. However, the life of an app on the Play Store depends on its success parameters. By accurately predicting the success of mHealth applications based on their features, app designers can foresee which apps are most likely to be installed and stay on the Play Store for a longer period.

Methods: Initially, a literature survey is conducted to investigate the existing methods/models of predicting app success and what parameters are used to assess their performance. Then, machine learning models appropriate for the problem at hand are identified. The apps related to health care and medicine are filtered from the entire dataset with 10 features, and experimental analysis is performed. To accurately predict the success factor, five Machine Learning classifier models are used on a publicly available dataset on the Kaggle platform. The models employed in this report are Logistic Regression, Gaussian Naïve Bayes, Decision Tree, SVC, and Gradient Boosting. The correlation matrix is utilized to investigate the

association among the attributes. The confusion matrix and comparison between the models using accuracy measures (Recall, Precision, Accuracy, and F1-score) helped to choose the best model to forecast the app's future and make wise decisions.

Results: The results findings imply that the Gradient Boosting algorithm outperforms with 88% accuracy in making predictions among all five machine learning classification algorithms. The results are best supported by using four different evaluation metrics: Precision, Recall, F1-score, and Accuracy. Finally, K-fold cross-validation also supported the Gradient Boosting algorithm with nearly 88% accuracy.

Conclusions: According to the result analysis, it can be inferred that all 5 machine learning classifiers used in the study were competent enough to predict success, but the Gradient Boosting algorithm surpassed in terms of accuracy.

Keywords: Mobile Applications, Prediction, Android, Success, Machine Learning, Health Care

1. Introduction

1.1 Background

Apps for mobile health, or mHealth apps, are growing in popularity as people's emphasis shifts towards their health. Based on projections made by Precedence Research, the global mHealth (mobile health) market will attain $340.5 billion by 2030 at a CAGR of 26.79% (Haldiya 2024). mHealth apps have transformed the manner in which health information is attained and upgraded the value of services in the healthcare field (Sadegh et al. 2018; Zhang et al. 2018). According to Zion Market Research (Research 2024), the market for mHealth apps is expected to reach a value of about USD 111 billion by 2025.

As the sooq for mobile devices has considerably improved in the last two decades, the mobile app development and deployment on respective app stores of the mobile device has also amplified. Unlike many other mobile OS, Android apps play a major role in the mobile software industry. The

Google Play Store on Android mobile devices provides access to millions of apps to work on. Daily, this platform is flooded with an enormous number of fresh apps. Developers of these apps are anticipating that their efforts won't be wasted and try to build an app that will be successful in this competitive market. It would be beneficial for the developers to know in advance if their app will be successful or not before uploading it to the Google Play Store. It would prove to be highly advantageous for developers to have a method that can help predict the success possibility of an application.

The success of an app can be ascertained by features like ratings, number of installs, and reviews, as opposed to the revenue it generates. Generally, many apps in the Play Store are not charged; the revenue generated by subscriptions, in-app purchases, and in-app advertisements is practically unknown. Although many apps are added to the Play Store daily, only a few apps achieve their monetary benefits and endure in their respective competitive marketplaces. It would be really helpful if the possibility of app success could be ascertained in advance.

This research paper aims to investigate a Machine Learning algorithm to predict the success of a healthcare app before it is uploaded to the Google Play Store. To achieve this objective, relevant features are explored that enable an app to succeed or fail. Many studies have been conducted to improve the prediction using various methods. Some of the identified literature is provided, which forms the background for this study.

In the paper by (Tuckerman 2014), the author used the mobile app data from the Google Play Store. From the data available, the author extracted the features of apps and then used a modeling approach to conclude that predictions can be made using features like no. of installs and user ratings. The author concluded that PCA with Linear regression presented the best pattern for prediction.

In the paper by (Guerrouj and Baysal 2016)The authors focused on API quality along with other features of the app for the success of the mobile app. They determine this quality based on bugs removed in APIs and what modifications were made in API methods.

Another approach by the authors in (Yao et al. 2017) considered versions released for the apps to recommend the app for its success. They addressed that the rating for an app, which is a major factor for the success of an app, highly changes when a new version of the same application is released.

Mueez et al. Khushba et al. (2018), considered a dataset obtained from the Google Play Store. The dataset obtained is used in predicting the success of an app. They based their prediction first on features like ratings, length of the app name, and the no. of installs for an app. Then they took the reviews under consideration to make predictions.

Suleman et al.(Suleman, M., Malik, A., & Hussan 2019), also used a Google Play Store dataset and then worked on seven features having 10839 apps in the dataset. They used machine learning techniques in predicting app success. The emphasis of their study concluded which model performed best in predicting app ratings.

In the paper by (Businge et al. 2019), the authors distinguished the apps based on their popularity of the apps. They collected data sets from two marketplaces, GitHub and the Google Play Store. Then they introduced social and technical factors to understand how they affect the popularity of an app and hence its success.

In the paper by Muradul Bashir et al. (2019), the focus of the authors for predicting the success of mobile apps on Google Play was based on user rating and the no. of installs of apps. The authors analyzed a repository of the Google Play Store and applied various machine learning algorithms. The authors later concluded that the SVM model generated the most accurate predictions.

In the study by Singh (2020), the authors operated on a dataset having Android apps and implemented data evaluation using machine learning methods. They then aimed to understand features/factors in apps that most impact their success.

In the research by Dehkordi et al. (2020), the authors presented a repository consisting of 100 successful and 100 unsuccessful apps from the Google Play Store. This repository consisted of 34 features that are used as input to

a neural network in predicting app success on the Play Store.

In the paper presented by Zhong et al. (2020), the authors collected data from the Google Play store, having 10,841 mobile apps with features such as app size, free or paid. Then the authors presented a model by using mobile app features and user reviews in the model to predict the possible rate.

The authors in Rathod et al. (2020) suggested a framework that predicts the success of mobile apps not only based on no. of installs, ratings, and reviews but also on the sentiments of customers. They have applied various machine learning algorithms to a dataset collected from the Google Play Store Data. Lastly, they incorporated the Voting Ensemble technique to improve the accuracy of algorithms.

In the study by Shashank and Naidu (2020), the authors intended to predict the success of an app based on the ratings of the app on the Google Play Store. The authors used machine learning Algorithms for predicting success. The dataset for analysis and prediction is collected from the Kaggle platform.

In a study by Prakash Reddy and Nallabolu (2020), the authors only presented explanatory analysis for a dataset from the Google Play store, having 2,67,000 apps. Their analysis can help app developers make better judgments on app reachability to their intended customers.

The authors of a study, Magar, Mali, and Abdelfattah (2021), presented the success prediction as a classification problem. They started by classifying the mobile apps based on numerous degrees of success. Based on this theory, the authors used various classification models to comprehend the app's success. The models are then assessed based on their performance rates.

In the study by (Tafesse 2021), the authors used a dataset obtained from the Apple App Store. The dataset consist apps from the gaming and productivity categories. The author examined how app ratings and user reviews influence the number of downloads of mobile applications. The findings revealed that both ratings and reviews have a direct impact on the number of downloads, indicating that higher ratings and more positive

reviews significantly increase the likelihood of an app being downloaded.

In the paper by Davazdahemami et al. (2023), predictive network analytics, deep learning, and artificial intelligence are used that focus on the developer-oriented aspect of an app. They worked on two aspects of the problem. One was focused on recommending developers to distinguish the suitable customers. The second work focused on providing perceptions in terms of design principles and factors for the app developers to explore the possibilities for app success.

In the study by Pattanaik, Priyadarshini (2023), the authors used a dataset from the Google Play Store to forecast the success of the app. Again, the authors extracted features of apps available in the dataset and employed machine learning techniques to analyze data from varied metrics and ascertain relationships.

Table 1.1 depicts a summarized comparison of various existing success predictions. It can be inferred that machine learning techniques are the most followed approach for prediction in the majority of the research papers. The factors that affect each method are also listed for prediction. Each method is evaluated on some parameter. Table 1.2 lists the prominent evaluation metrics followed in existing studies.

Table 1.1 Comparison of existing success prediction approaches

References	Approach	Factors	Machine Learning	Method
(Guerrouj and Baysal 2016)	Model-Based	App size, category, reviews, rating	No	Multiple Linear Regression
(Yao et al. 2017)	Algorithmic	Ratings, reviews, versions	No	Version-Aware rating prediction
(Khushba et al. 2018)	Model-Based	App name, category, content rating, installs, rating, last updated, type, rating text, review text, size	Yes	Random Forest, XBoost Classifier, K-NN, SVM
(Suleman, M., Malik, A., & Hussan 2019)	Model-Based	Review, size, install, type, content rating, version, rating	Yes	Linear Regression, Decision Tree, Logistic Regression, Naïve Bayesian, K-Means Clustering, KNN, ANN
(Businge et al. 2019)	Model-Based	User app downloads, avg rating, activities forked on GitHub	No	Multiple Linear Regression
(Muradul Bashir et al. 2019)	Model-Based	Installation, no, user rating, category, size, price	Yes	KNN, SVM, Random Forest
(Ruhel, S. 2019)	Algorithmic	Review, rating, category, size, install, type, content, genre, last updated	Yes	Backpropagation (ANN), Deep Learning (HIVE)
(Singh 2020)	Model-Based	Content rating, installs	Yes	SVM, random forest, Linear regression

(Dehkordi et al. 2020)	Algorithmic	Name, No. of packages, No. Of classes, No. Of Methods, LOC, size	Yes	MLP, SVM, Random Forest, Decision trees, ANN, LVQ, NPR, PCA
(Zhong et al. 2020)	Model-Based	Category, reviews, size, installs, type, price, content rating, android version	Yes	Random forest, MLP, gradient boosting, Logistic regression
(Rathod et al. 2020)	Model-Based	Installs, rating, review, type, size, price, genre, user group, last updated, android version	Yes	KVM, SVM, Decision trees, Random Forest
(Shashank and Naidu 2020)	Model-Based	App category, rating, review, installs, size, type, price, content rating, genre	Yes	KNN, SVM, Linear regression, Random Forest, K- K-means clustering
(Prakash Reddy and Nallabolu 2020)	NA	App category, rating, review, installs, size, price, paid/free	No	N.A., only analysis
(Magar, Mali, and Abdelfattah 2021)	Model-Based	rating, installs, size, price, type	Yes	KNN, SVM, logistic regression, Random Forest, Decision trees, SGD
(Tafesse 2021)	Model-Based	App rating, size, App version (major), App version. (minor), Revenue, app name, app value, Download, reviews, update Regency	Yes	regression, Random Forest
(Pattanaik, Priyadarshini 2023)	Model-Based	Rating, reviews, size, installs, type, price, content rating, genre, last update	Yes	Random Forest, Decision trees, SVM, SGB, GRU

Table 1.2. Evaluation Metrics used by different studies

Evaluation metrics	Various Studies
R, R^2	(Guerrouj and Baysal 2016), (Suleman, M., Malik, A., & Hussan 2019), (Businge et al. 2019), (Singh 2020), (Shashank and Naidu 2020), (Tafesse 2021)
RMSE, MAE	(Yao et al. 2017), (Suleman, M., Malik, A., & Hussan 2019)
MSE	(Khushba et al. 2018),(Suleman, M., Malik, A., & Hussan 2019), (Muradul Bashir et al. 2019),(Singh 2020), (Dehkordi et al. 2020),(Zhong et al. 2020), (Rathod et al. 2020),(Shashank and Naidu 2020), (Tafesse 2021)
Cost function	(Ruhel, S. 2019)
Precision, recall, F-score, Accuracy	(Magar, Mali, and Abdelfattah 2021), (Pattanaik, Priyadarshini 2023)

2. Methods

Fig. 1.1 depicts the flow of the method adopted for conducting experimental work using Python Jupyter Notebook 7.0.8. Below is a breakdown of each stage.

1. Data Collection

In this research work, the mobile app dataset was collected from Kaggle.com (Jithin 2021), which is a hub for data scientists to publish and share their work. Mobile apps related to the health care and medicine category of the Google Play app store dataset are considered in this study. The variables such as App Name, Category, LOC, APK Size, Avg Rating, Installs, Version, Free/Paid, Age group, Genres, Last Updated, Current Version, Min required Android version form the columnar part of the dataset.

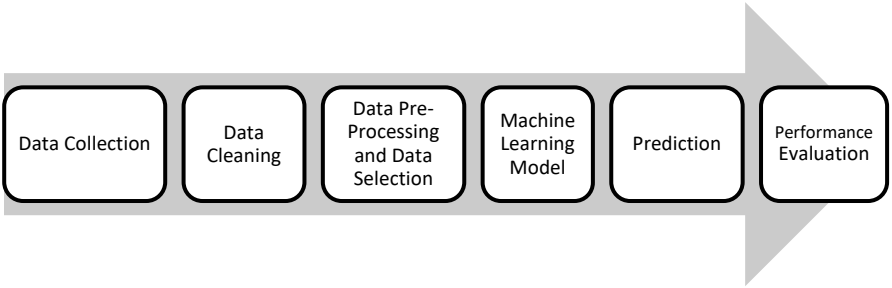


Fig 1.1 Methodology Adopted

2. Data Cleaning

The dataset initially had 10841 rows (mobile apps) and 13 columns (features of the app). After selecting health and medicine-related apps, only 804 mobile apps were left. The duplicate, missing, and inappropriate rows were removed and left with 564 rows. Relevant features or attributes required for success prediction are chosen. Fig. 1.2 gives some insights into the final dataset. There are 564 apps of the 2 categories (Health and medical), with ratings from 1 to 5, and the Calorie Counter-MyFitnessPal App is at the top of the medical category. Fig. 1.3 shows the total apps in each category, with nearly 300 apps under the medical category.

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
count	564	564	564.000000	5.640000e+02	564	5.640000e+02	564	564	564	564	564	564	564
unique	534	2	NaN	NaN	160	NaN	2	30	4	2	257	315	21
top	Calorie Counter - MyFitnessPal	MEDICAL	NaN	NaN	Varies with device	NaN	Free	0	Everyone	Medical	27-Jul-18	Varies with device	4.1 and up
freq	3	302	NaN	NaN	88	NaN	489	489	523	302	20	81	118
mean	NaN	NaN	4.219149	5.716577e+04	NaN	2.487888e+06	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	0.668491	2.590531e+05	NaN	2.185135e+07	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	1.000000	1.000000e+00	NaN	1.000000e+00	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	4.000000	3.375000e+01	NaN	1.000000e+03	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	4.400000	5.745000e+02	NaN	5.000000e+04	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	4.600000	1.557725e+04	NaN	1.000000e+06	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	5.000000	4.559407e+06	NaN	5.000000e+08	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 1.2 Description of the dataset after selecting Health care and Medical Apps

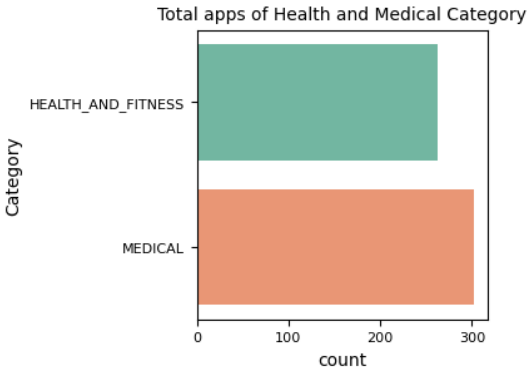


Fig. 1.3 Total Apps of Health Care and Medical category

3. Data Pre-processing and Data Selection

Data is prepared for classification to enhance classifier accuracy, and *the Installs* column is chosen as the target variable for success prediction. The following are the preprocessing tasks performed on the data to prepare for classification models: -

1. Casting of *Installs* to int, *Avg. Rating* and P*rice* to float (and remove $ from Price column).
2. Encode features (*Type, Category*) by the label encoding method.
3. Replaced *the Size* column with unified size values by converting Megabytes and Kilobytes into Bytes, then normalizing App size values.
4. Remove unimportant features like (*App Name, Versions, Last updated*).

The Installs column is considered as targets/labels for classifier Models. As App success is based on the *Installs*, there is a need to categorize the target as follows

1. If the number of *Installs* is less than 5000, then it's a failed app and is coded as 0.
2. If the number of *Installs* is greater than 5000 and less than 1,000,000, then the limited success *app,* is coded as 1.

3. If the number of *Installs* is greater than 1,000,000 then a successful app, is coded as 2.

Fig. 1.4 shows the final data set after pre-processing, encoding, and categorizing *the Installs* column. The dataset is also checked to see if it is equally distributed under all three categories of *the Installs* column. The result shows that 30% fall in the 0 (failed app) category, 41% for 1(limited success app) category, and 29% for 2 (successful app).

	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres_Health & Fitness	Genres_Medical
669	0	5.0	4	0.139998	0	0	0.00	0	1	0
638	0	3.9	26	0.119998	0	0	0.00	0	1	0
247	1	3.0	4	0.033998	0	1	0.99	0	0	1
608	0	4.4	622	0.024998	1	0	0.00	0	1	0
611	0	5.0	3	0.092998	0	0	0.00	0	1	0

Fig. 1.4 Dataset after coding of the Installs variable in 3 categories (0,1,2) (success predictor)

4. Machine Learning Models

Five Machine Learning models are carefully chosen as a classification problem (success or fail). The selected ones are as follows: -

1. *Logistic Regression Classifier* (Hancock et al. 2023)*:* LRC is applied in binary classification where the sigmoid function is used, taking independent variables as input and producing a probability value between 0 and 1.
2. *Naive Bayes Classifier* (Shaban et al. 2021)*:* NBC is based on Bayes' Theorem, which defines the probability of a certain event, built on the aforementioned knowledge of conditions that could be associated with this event.
3. *Decision Tree Classifier* (Singh 2020), DTC's purpose is to generate a model that predicts an objective built on input attributes using Tree models. The leaves in models signify class labels, and feature conjunctions are represented by branches.

4. *Support Vector Machine classifier* (Hancock et al. 2023) SVM is considered a robust classifier that aims to find an optimal hyperplane that divides feature space by finding how far the margin is from the hyperplane, and the traits that are closest across all classes.
5. *Gradient Boosting Classifier* (Hancock and Khoshgoftaar 2021), GBC is a strong boosting algorithm that unites numerous weak learners to robust learners, in which each new model is trained to minimize the loss function.

5. Prediction

The Machine Learning models are used to generate/fit the model and predict the values after splitting the dataset into training (75%) and testing (25%). The time (in seconds) is also recorded for each classifier to compare the performance metric.

6. Performance Evaluation

For performance evaluation, four different measures: Precision, Recall, F1-score, and Accuracy (Sohil, Sohali, and Shabbir 2022; Bognár and Fauszt 2022) are used. The ratio of True Positives to all Positives is known as precision. The recall formula is calculated by dividing the total number of positive results and false negative results. The F-score is a measure of predictive performance ranging from 0 to 1.

The ratio of all correctly identified observations to total observations, which ranges from 0 to 1, is known as Accuracy (Sohil, Sohali, and Shabbir 2022). The k-fold Cross-validation method is used to test the classifier's reliability. Cross-validation divides the training data into k parts (k=10 in this implementation), then uses 9 parts as a training dataset and tests with the remaining part (validation dataset). This operation is repeated 10 times, and then the final result is the average score of 10 trials.

3. Results

In this study, five classification models of the dataset are subjected to machine learning. The 10 features selected for fitting the models are examined for further processing using a heatmap. The correlation between

all features is assessed to check if they are independent of each other. Fig. 1.5 shows that there is very little correlation between all the features.

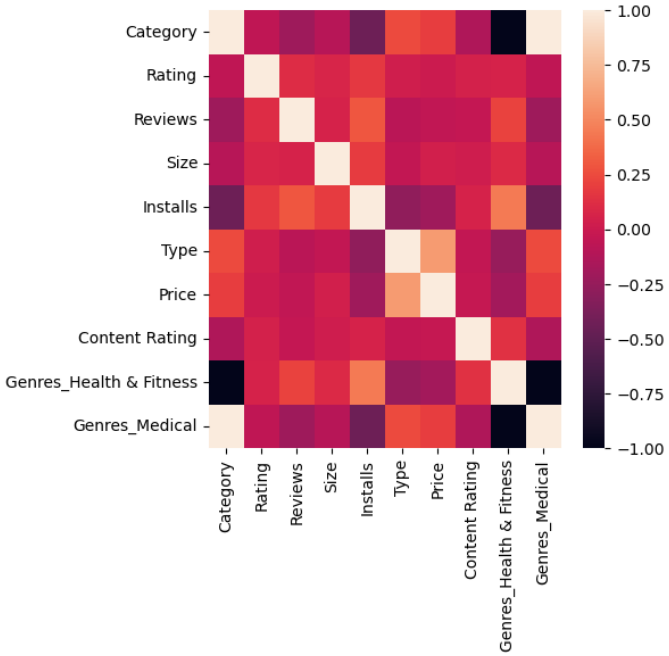


Fig. 1.5 Correlation between features using Heatmap

Following model fitting on the final dataset, each model is assessed using the classification report and confusion matrix produced on four distinct metrics: accuracy, precision, recall, and F1-score. The confusion matrix shows true positives, false positives, true negatives, and false negatives in a visual manner in Fig. 1.6. When combined, these metrics provide useful data regarding the model's overall effectiveness, aiding in the assessment and improvement of its classification capabilities.

The results in Table 1.3 and Fig.1.7 for F1-scores show that prediction accuracy above 80% is for Gradient Boosting, Gaussian Naïve Bayes, and Decision Tree. The best prediction accuracy is given by Gradient Boosting with 88%. Performance-wise, Decision Tree succeeded with 0.0094 secs,