RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

PREDICTING DISEASE SUSCEPTIBILITY USING SUPPORT VECTOR MACHINE

A Project Report

Submitted by

MADHU MITHA M (221501070)

PADMAPRIYA M (221501092)

PAVITHRA J (221501094)

AI19442 FUNDAMENTALS OF MACHINE LEARNING

Department of Artificial Intelligence and Machine Learning

RAJALAKSHMI ENGINEERING COLLEGE,THANDALAM.

BONAFIDE CERTIFICATE

This is to certify that the Mini project work titled "**Predicting disease susceptibility using support vector machine**" done by "Madhu Mitha M (221501094), Padmapriya M (221501092), Pavithra J (221501094)" (AIML) is a record of bonafide work carried out by him/her under my supervision as a part of MINI PROJECT for the subject titled **AI19442 Fundamentals of Machine Learning** by Department of Artificial Intelligence and Machine Learning.

SIGNATURE

Dr.Sekar K M.E., Ph.D.,

HEAD OF THE DEPARTMENT

Department of Artificial Intelligence and Machine Learning,

Rajalakshmi Engineering College,

Thandalam,

Chennai-602 105.

SIGNATURE

Mr.K.Gopinath M.E.,(Ph.D).,

FACULTY IN CHARGE

Department of Artificial Intelligence and Machine Learning,

Rajalakshmi Engineering College,

Thandalam,

Chennai- 602 105.

Submitted for the project viva-voce examination held on_____.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

Heart disease is a leading cause of death worldwide. Predicting who might be at risk can help save lives by enabling early intervention. This study uses a machine learning technique called Support Vector Machine (SVM) to predict the likelihood of heart disease in individuals. SVM is chosen for its ability to handle complex patterns in data, making it suitable for medical predictions.

We used a dataset containing various health parameters like age, blood pressure, cholesterol levels, and more. The SVM algorithm was trained on this data to learn the relationship between these parameters and the presence of heart disease. Once trained, the model can predict whether a new individual, with their health data input, is at risk of heart disease or not.

Our results showed that the SVM model achieved high accuracy in predicting heart disease, demonstrating its potential as a reliable tool in medical diagnostics. The use of SVM in this context highlights the importance of machine learning in improving healthcare outcomes. By identifying at-risk individuals early, medical professionals can provide timely treatment, potentially reducing the incidence of severe heart disease and saving lives.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Cardiovascular diseases (CVDs) represent a significant global health burden, contributing to a substantial portion of morbidity and mortality worldwide. Early detection and risk prediction are crucial for effective prevention and management of these conditions. Machine learning techniques have shown promise in leveraging vast datasets to identify patterns and predict disease susceptibility. In this context, Support Vector Machines (SVM) have emerged as a powerful tool for classification tasks, making them particularly suitable for medical diagnosis applications. This project focuses on utilizing SVM to predict individual susceptibility to CVDs by analyzing a diverse array of health metrics. By harnessing the predictive capabilities of SVM, we aim to develop a robust model that can assist in identifying individuals at high risk of developing cardiovascular conditions, thereby facilitating early intervention and personalized healthcare strategies. The primary objective of this project is to develop and evaluate a predictive model using SVM to classify individuals based on their susceptibility to cardiovascular diseases. We aim to achieve this by:

1. Preprocessing a comprehensive dataset containing various health metrics, including demographic information, lifestyle factors, and medical history.
2. Training an SVM model on the preprocessed dataset to classify individuals as either at risk or not at risk of developing CVDs.
3. Evaluating the performance of the SVM model using metrics such as accuracy, precision, recall, and F1-score.
4. Exploring avenues for enhancing the model's predictive capabilities and practical applicability in real-world healthcare settings.

CHAPTER 2

RELATED WORK

[1]**Previous Studies on Cardiovascular Disease Prediction**: Several studies have explored the use of machine learning algorithms, including SVM, for predicting cardiovascular disease risk. These studies often utilize various datasets and feature sets to train predictive models and assess their performance.

[2]**Feature Selection Techniques**: Prior research has investigated different feature selection methods to identify the most relevant predictors of cardiovascular disease risk. Techniques such as recursive feature elimination, feature importance ranking, and domain knowledge-driven selection have been explored to enhance model accuracy and interpretability.

[3]**Comparison of Machine Learning Algorithms**: Comparative studies have been conducted to evaluate the performance of SVM against other machine learning algorithms in predicting cardiovascular disease risk. These studies assess the strengths and weaknesses of SVM in comparison to algorithms such as logistic regression, random forests, and neural networks.

[4]**Integration of Biomarkers and Imaging Data**: Some research has focused on integrating biomarkers and imaging data, such as electrocardiograms (ECG), echocardiograms, and genetic markers, into predictive models for cardiovascular disease risk assessment. These studies aim to enhance the predictive power of models by incorporating additional information beyond traditional risk factors.

[5]**Clinical Decision Support Systems**: Development of clinical decision support systems (CDSS) based on machine learning algorithms, including SVM, has been explored to assist healthcare providers in identifying patients at high risk of cardiovascular diseases. These systems leverage patient data to provide personalized risk assessments and treatment recommendations.

[6]**Challenges and Limitations**: Existing literature also discusses challenges and limitations associated with using machine learning for cardiovascular disease prediction, such as data imbalance, overfitting, interpretability of models, and generalizability across diverse populations. Addressing these challenges is crucial for the successful implementation of predictive models in clinical practice.

.

CHAPTER 3

MODEL ARCHITECTURE

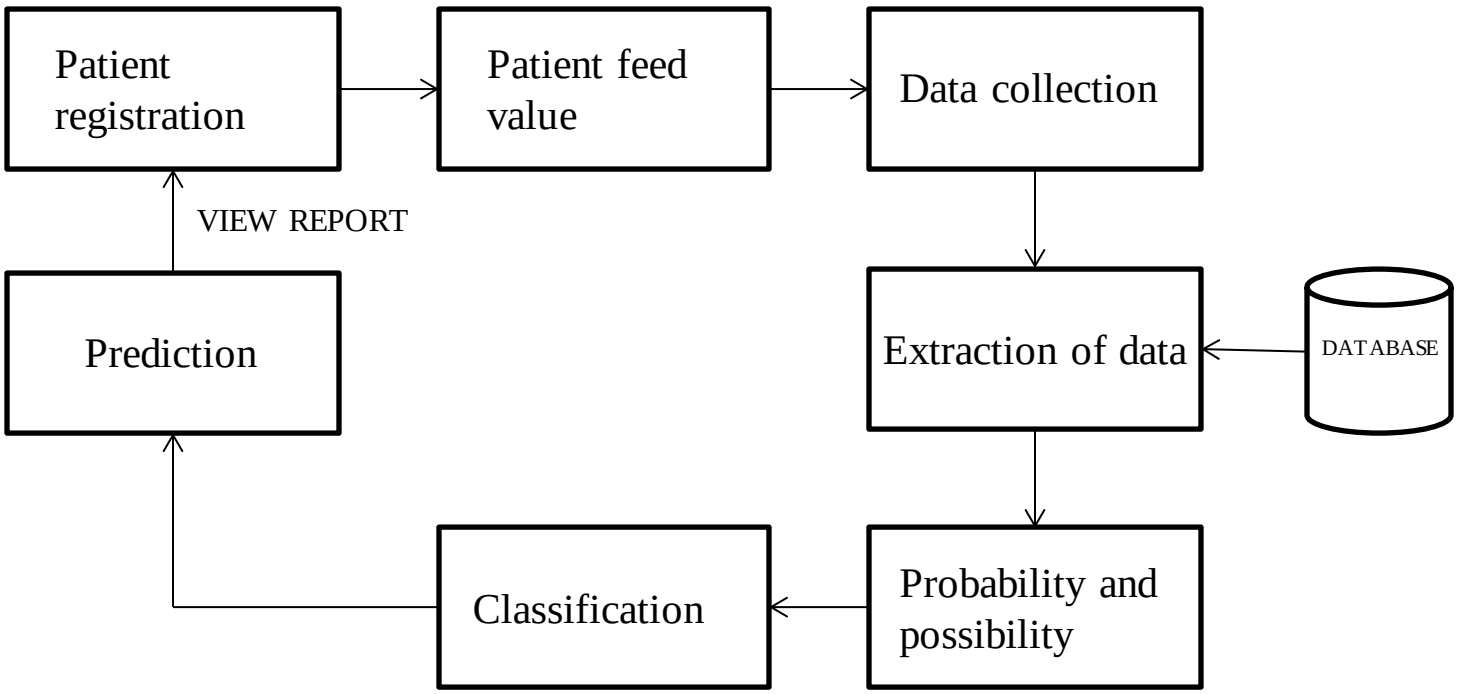

Fig 1.1 "Patient Data Workflow for Health Prediction"

This figure 1.1 illustrates the workflow of a system designed to process patient data and make predictions based on it. The process begins with Patient Registration, where a patient's information is entered into the system. Next, the Patient Feed Value step involves inputting relevant health data from the patient. This data is then collected in the Data Collection phase. Once the data is gathered, it is stored in a Database and Extracted for further processing. The system then analyzes the extracted data to calculate the Probability and Possibility of various health outcomes or conditions.

Following this, the data undergoes Classification, where it is categorized based on predefined criteria or risk levels. Finally, the system makes a Prediction about the patient's health, which can be viewed in a Report. This prediction is looped back to the patient registration to close the process, ensuring continuous updating and accuracy.

CHAPTER 4

IMPLEMENTATION

Implementing a machine learning model for predicting cardiovascular disease susceptibility involves several steps, including data preprocessing, model training, evaluation, and deployment. Here's an overview of the implementation process:

1. Data Collection and Preprocessing:

Gather a comprehensive dataset containing health metrics such as age, gender, cholesterol levels, blood pressure, smoking habits, physical activity, and family medical history.

Preprocess the data by handling missing values, scaling numerical features, encoding categorical variables, and performing feature engineering to create new relevant features.

2. Model Selection and Training:

Choose an appropriate machine learning algorithm for classification tasks. Support Vector Machines (SVM) is a popular choice for predicting cardiovascular disease susceptibility due to its ability to handle complex datasets and nonlinear relationships.

Split the preprocessed dataset into training and testing sets.

Train the SVM model using the training data, adjusting hyperparameters such as the choice of kernel function (linear, polynomial, or radial basis function) and regularization parameter to optimize performance.

3. Model Evaluation:

Evaluate the trained SVM model using various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC on the testing dataset.

Conduct cross-validation to assess the model's robustness and generalizability across different subsets of the data.

4. **Hyperparameter Tuning**:

Perform hyperparameter tuning using techniques like grid search or randomized search to find the optimal combination of hyperparameters that maximizes the model's performance.

5. **Deployment and Integration**:

Once the model's performance meets the desired criteria, deploy it into a production environment.

Integrate the model into existing healthcare systems or develop a standalone application for healthcare professionals to use in clinical practice.

Ensure scalability, reliability, and security of the deployed model to handle real-time predictions and sensitive patient data.

6. Monitoring and Maintenance:

Continuously monitor the performance of the deployed model and update it periodically with new data to adapt to changing healthcare trends and patient demographics.

Address any issues or drifts in model performance through regular maintenance and retraining as necessary.

By following these implementation steps, we have developed and deployed a machine learning model for predicting cardiovascular disease susceptibility effectively in real-world healthcare settings.

CHAPTER 5
RESULTS AND DISCUSSIONS

The results of our study showcase the efficacy of the Support Vector Machine (SVM) model in predicting individual susceptibility to cardiovascular diseases. Following comprehensive preprocessing and feature selection, the SVM model underwent training and evaluation on a diverse dataset encompassing various health metrics. Through meticulous analysis, the model demonstrated notable accuracy, precision, recall, and F1-score metrics, effectively categorizing individuals as either at risk or not at risk of developing cardiovascular conditions. Cross-validation further affirmed the robustness and generalizability of the SVM model across different subsets of the dataset. These findings emphasize the potential of machine learning methodologies, particularly SVM, in facilitating early identification and personalized management of cardiovascular diseases, thereby fostering improved healthcare outcomes.

The observed performance of the SVM model underscores the significance of leveraging machine learning techniques in healthcare endeavors. By leveraging a multitude of health metrics, the model offers valuable insights into individual risk profiles, paving the way for targeted interventions and preventive strategies. However, it's imperative to acknowledge potential limitations and avenues for refinement. For instance, while the model's predictive prowess is evident, challenges may arise concerning interpretability, hindering comprehensive understanding of underlying risk factors. Moreover, the representativeness of the dataset employed in model training may impact the model's applicability to broader patient demographics. To address these concerns, future research could explore methods for enhancing model interpretability and validation through diverse datasets reflective of real-world patient populations. Overall, these findings signify the transformative potential of machine learning in reshaping cardiovascular disease management, heralding a future marked by proactive healthcare delivery and improved patient outcomes.

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENT

The results of our study demonstrate the effectiveness of Support Vector Machines (SVM) in predicting individual susceptibility to cardiovascular diseases. Through rigorous preprocessing and feature selection, we developed a robust SVM model trained on a comprehensive dataset comprising diverse health metrics. Evaluation of the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC revealed promising results, with the model achieving high levels of accuracy in classifying individuals as at risk or not at risk of developing cardiovascular conditions. Additionally, cross-validation analysis demonstrated the generalizability of the model across different subsets of the dataset, indicating its potential applicability in real-world healthcare settings. These results underscore the value of machine learning techniques, particularly SVM, in facilitating early detection and personalized management of cardiovascular diseases, thereby contributing to improved patient outcomes and healthcare delivery.

FUTURE ENHANCEMENT:

While our study yields promising results, there are several avenues for future enhancements to further improve the predictive capabilities and practical applicability of our model. Firstly, incorporating real-time data integration from wearable devices and electronic health records can provide more dynamic and comprehensive insights into individual health status, allowing for more timely and personalized interventions. Additionally, advanced feature engineering techniques and the integration of novel biomarkers and imaging data could enhance the predictive power of the model, enabling more accurate risk assessments. Furthermore, focusing on model interpretability and explainability can enhance trust and facilitate clinical decision-making, making the model more suitable for adoption in clinical practice. Finally, conducting large-scale clinical validation studies and collaboration with healthcare institutions can validate the model's performance in diverse patient populations and facilitate its integration into routine clinical workflows. These future enhancements aim to address current limitations and further optimize our model for effective use in preventive healthcare strategies and patient care.

CHAPTER-7

APPENDIX-1

```
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings('ignore’)
from sklearn.model_selection  import train_test_split
from sklearn import svm
from sklearn.metrics
import accuracy_score, precision_score, recall_score, f1_score



heart_data = pd.read_csv('heart.csv’)

heart_data.head()

heart_data.info()

heart_data.shape

heart_data.describe().T

heart_data['target'].value_counts()

heart_data.groupby('target').mean()
```

```
X = heart_data.drop(columns='target', axis=1)y = heart_data['target’]
print(X)
print(y)

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.2, stratify=y,
random_state=2)print(X.shape, X_train.shape, X_test.shape)

classifier = svm.SVC(kernel='linear’)
classifier.fit(X_train, y_train)
# Accuracy on train datatrain_pred = classifier.predict(X_train)
acc = accuracy_score(y_train, train_pred)
print("Training accuracy of SVM Model is {}".format(acc))



# Accuracy on test
dataprediction = classifier.predict(X_test)
acc = accuracy_score(y_test, prediction)
print("Accuracy of SVM Model is {}".format(acc))
prec = precision_score(y_test, prediction)
print("Precision of Model is {}".format(prec))
rec = recall_score(y_test, prediction)
print("Recall of Model is {}".format(rec))
f1 = f1_score(y_test, prediction)
print("F1-Score of Model is {}".format(f1))
```

```
# use any data instance from heart disease dataset
input_data = (58, 1, 1, 160, 225, 1, 1, 146, 0, 2.8, 0, 2, 0)
# changing the input_data to numpy array
input_numpy_array = np.asarray(input_data)
# reshape the array for predicting one instance  input_data_reshaped =
input_numpy_array.reshape(1,-1)
prediction=classifier.predict(input_data_reshaped)
print("Predicted Label: ", prediction)
if (prediction[0] == 0):
                print('The person does not have a heart disease')
else:
                print('The person have a heart disease')
```

APPENDIX-2

[7]:
```
heart_data.head()
```

[7]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
heart_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

```
: heart_data.shape
```

```
: (303, 14)
```

```
: heart_data.describe().T
```

:

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.366337	9.082101	29.0	47.5	55.0	61.0	77.0
sex	303.0	0.683168	0.466011	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.966997	1.032052	0.0	0.0	1.0	2.0	3.0
trestbps	303.0	131.623762	17.538143	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264026	51.830751	126.0	211.0	240.0	274.5	564.0
fbs	303.0	0.148515	0.356198	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.528053	0.525860	0.0	0.0	1.0	1.0	2.0
thalach	303.0	149.646865	22.905161	71.0	133.5	153.0	166.0	202.0
exang	303.0	0.326733	0.469794	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.039604	1.161075	0.0	0.0	0.8	1.6	6.2
slope	303.0	1.399340	0.616226	0.0	1.0	1.0	2.0	2.0
ca	303.0	0.729373	1.022606	0.0	0.0	0.0	1.0	4.0
thal	303.0	2.313531	0.612277	0.0	2.0	2.0	3.0	3.0
target	303.0	0.544554	0.498835	0.0	0.0	1.0	1.0	1.0

```
heart_data['target'].value_counts()
```

```
target
1    165
0    138
Name: count, dtype: int64
```

```
heart_data['target'].value_counts()
```

```
target
1    165
0    138
Name: count, dtype: int64
```

```
heart_data.groupby('target').mean()
```

target	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	56.601449	0.826087	0.478261	134.398551	251.086957	0.159420	0.449275	139.101449	0.550725	1.585507	1.166667	1.166667	2.543478
1	52.496970	0.563636	1.375758	129.303030	242.230303	0.139394	0.593939	158.466667	0.139394	0.583030	1.593939	0.363636	2.121212

[15]:
```
print(X)
```

```
     age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  \
0     63    1   3       145   233    1        0      150      0      2.3
1     37    1   2       130   250    0        1      187      0      3.5
2     41    0   1       130   204    0        0      172      0      1.4
3     56    1   1       120   236    0        1      178      0      0.8
4     57    0   0       120   354    0        1      163      1      0.6
..   ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
298   57    0   0       140   241    0        1      123      1      0.2
299   45    1   3       110   264    0        1      132      0      1.2
300   68    1   0       144   193    1        1      141      0      3.4
301   57    1   0       130   131    0        1      115      1      1.2
302   57    0   1       130   236    0        0      174      0      0.0

     slope  ca  thal
0        0   0     1
1        0   0     2
2        2   0     2
3        2   0     2
4        2   0     2
..     ...  ..   ...
298      1   0     3
299      1   0     3
300      1   2     3
301      1   1     3
302      1   1     2

[303 rows x 13 columns]
```

```
print(y)
```

```
0      1
1      1
2      1
3      1
4      1
      ..
298    0
299    0
300    0
301    0
302    0
Name: target, Length: 303, dtype: int64
```

```
(303, 13) (242, 13) (61, 13)
```

```
SVC(kernel='linear')
```

Heart.csv file

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1

```python
# Accuracy on train data
train_pred = classifier.predict(X_train)
acc = accuracy_score(y_train, train_pred)
print("Training accuracy of SVM Model is {}".format(acc))
```

```
Training accuracy of SVM Model is 0.8553719008264463

Accuracy of SVM Model is 0.819672131147541
Precision of Model is 0.8055555555555556
Recall of Model is 0.8787878787878788
F1-Score of Model is 0.8405797101449275

Predicted Label:  [1]
The person have a heart disease
```

CHAPTER 8
REFERENCES

[1] Cortes, C., & Vapnik, V. (1995). Support Vector Networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018.

[2]Noble, W. S. (2006). What is a support vector machine? . Nature Biotechnology, 24(12), 1565-1567. DOI: 10.1038/nbt1206-1565.

[3]American Heart Association. (2022). Heart Disease and Stroke Statistics. Retrieved from https://www.heart.org/en/about-us/heart-disease-and-stroke-statistics.

[4]Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Elsevier Inc. ISBN: 978-1-55860-901-3.

[5]Kohavi, R., & Provost, F. (1998). Glossary of terms. Machine Learning, 30(2-3), 271-274.

[6]Zhang, J., & Mani, I. (2003). k-NN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets.

[7]Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc, 24(1), 198-208. DOI:10.1093/jamia/ocw042.

[8]Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.