VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590018, KARNATAKA

An Internship Report
on

Breast Cancer Prediction

Submitted in partial fulfillment of the requirements during VIII Semester for the degree of **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

Submitted by

VIBHA S NAVALE
USN: 1RN17IS115

Under the Guidance of

Mrs. Hema N
Assistant Professor
Department of ISE

Department of Information Science and Engineering

RNS Institute of Technology

Dr. Vishnuvardhan Road, Rajarajeshwari Nagar post
Channasandra, Bengaluru-560098

2020 – 2021

RNS INSTITUTE OF TECHNOLOGY

Dr. Vishnuvardhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru - 560098

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled ***Breast Cancer Prediction*** has been successfully completed by **Vibha S Navale (1RN17IS115)**, a bonafide student of **RNS Institute of Technology, Bengaluru** in partial fulfillment of the requirements for the award of degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during academic year **2020-2021**. The internship report has been approved as it satisfies the academic requirements in respect of internship work for the said degree.

Mrs. Hema N
Internship Guide
Department of ISE

Dr. S Satish Kumar
Professor and HOD
Department of ISE

Dr. M K Venkatesha
Principal
RNSIT

External Viva

Name of the Examiners **Signature with Date**

1. ______________________ 1.______________________

2. ______________________ 2.______________________

DECLARATION

I, **VIBHA S NAVALE [USN: 1RN17IS115],** student of VIII Semester BE, in Information Science and Engineering, RNS Institute of Technology, hereby declare that the internship entitled ***Machine Learning: Algorithms in the Real World*** has been carried out by me and submitted in partial fulfillment of the requirements during *VIII Semester for the degree of* ***Bachelor of Engineering in Information Science and Engineering*** *of Visvesvaraya Technological University, Belagavi* during academic year 2020-2021.

Place: Bengaluru

Date:

VIBHA S NAVALE (1RN17IS115)

ABSTRACT

Breast cancer has been identified as the second leading cause of death among women worldwide after lung cancer and hence, it becomes extremely crucial to identify it at an early stage, which can considerably increase the chances of survival. The most important part in cancer detection is to be able to differentiate between benign and malignant tumors and this is where the work of Machine Learning comes in. Taking all the dependent features upon consideration, Supervised Machine Learning methods allow for classification with higher degree of accuracy and improve upon the misdiagnosis of the physicians, which might occur almost 20% of the time.

In this report, the project focuses towards understanding the shortcomings of digital mammograms in detection of breast cancer and utilize Machine Learning classifiers for the classification of benign and malignant tumors using the Wisconsin dataset. Apart from this, Supervised Machine Learning classifiers are considered for implementation, such as Logistic Regression, Decision Tree, Random Forest, K Nearest Neighbour (KNN), Support Vector Machine (SVM) and Naive Bayes classifiers for assessing the risks involved with breast cancer by analyzing the biomarkers that are involved with it. The aim is to provide a comprehensive view on prediction of breast cancer through Machine Learning through data analyses, which can play a pivotal role in prevention of misdiagnosis in future.

ACKNOWLEDGMENT

At the very onset I would like to place my gratefulness to all those people who helped me in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is they who guided me in the right direction.

First of all, I would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, I express sincere gratitude to our Director **Dr. H N Shivashankar** and the Principal **Dr. M K Venkatesha,** for providing us all the facilities.

I am extremely grateful to our own and beloved Professor and Head of Department of Information science and Engineering, **Dr. S Satish Kumar**, for having accepted to patronize me in the right direction with all his wisdom.

I place my heartfelt thanks to **Mrs. Hema N**, Assistant Professor, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

I also thank the internship coordinator **Mr. R Rajkumar,** Assistant Professor, Department of Information Science and Engineering. I would thank my friends for having supported me with all their strength and might. Last but not the least, I thank my parents for supporting and encouraging me throughout. I have made an honest effort in this assignment.

VIBHA S NAVALE
1RN17IS115

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Artificial intelligence

Artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals. In computer science AI research is defined as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving”. Modern machine capabilities generally classified as AI include successfully understanding human speech, competing at the highest level in strategic game systems (such as chess and Go), autonomously operating cars, intelligent routing etc.

Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism, followed by disappointment and the loss of funding (known as an "AI winter"), followed by new approaches, success and renewed funding. For most of its history, AI research has been divided into subfields that often fail to communicate with each other. These sub-fields are based on technical considerations, such as particular goals (e.g. "robotics" or "machine learning"), the use of particular tools ("logic" or artificial neural networks), or deep philosophical differences. Subfields have also been based on social factors (particular institutions or the work of particular researchers).

The traditional problems (or goals) of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. General intelligence is among the field's long-term goals. Approaches include statistical methods, computational intelligence, and traditional symbolic AI. Many tools are used in AI, including versions of search and mathematical optimization, artificial neural networks, and methods based on statistics, probability and economics. The AI field draws upon computer science, mathematics, psychology, linguistics, philosophy and many others.

The field was founded on the claim that human intelligence "can be so precisely described that a machine can be made to simulate it". This raises philosophical arguments about the nature of the mind and the ethics of creating artificial beings endowed with human-like intelligence which are issues that have been explored by myth, fiction and philosophy since antiquity. Some people also consider AI to be a danger to humanity if it progresses unabated. Others believe that AI, unlike previous technological revolutions, will create a risk of mass unemployment.

In the twenty-first century, AI techniques have experienced a resurgence following concurrent advances in computer power, large amounts of data, and theoretical understanding. AI techniques have become an essential part of the technology industry, helping to solve many challenging problems in computer science, software engineering and operations research. In simple terms, AI aims to extend and augment the capacity and efficiency of mankind in tasks of remaking nature and governing the society through intelligent machines, with the final goal of realizing a society where people and machines coexist harmoniously together.

1.2 Machine Learning

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

The name machine learning was coined in 1959 by Arthur Samuel. Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.

Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making using computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies.

1.3 Breast Cancer Prediction

Breast cancer became the major source of mortality among women. The accessibility of healthcare datasets and data analysis promote researchers to apply study in extracting unknown pattern from healthcare datasets. Hence it is important to predict the incidence of breast cancer at an early stage by analysing the smallest set of attributes selected from clinical dataset. And this can be achieved by using machine learning techniques.

Cancer starts when the cells begin to grow out of control. Breast cancer cells usually form a tumor that can often be seen on an X-ray or felt as a lump. Breast cancer occurs almost entirely in women, but sometimes men can get this too.

Most of the lumps are benign and not cancer (malignant). Non-cancerous breast tumors are abnormal growths, but they do not spread outside the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a health care professional to determine if it is benign or malignant (cancer) and if it might affect future cancer risk. Therefore, predicting if the tumor is benign or malignant becomes necessary and Machine Learning is helpful in such scenarios.

1.3.1 Objectives of the project:

1. To analyse and observe which features are most helpful in predicting benign or malignant cancer.
2. To see general trends that may aid in the model selection and avoid multicollinearity in the data.
3. To classify whether the breast cancer is benign or malignant and take necessary actions based on this outcome.
4. To find the best algorithm to achieve maximum accuracy for the prediction.

Chapter 2

LITERATURE REVIEW

Classification is the way toward finding a model (or capacity) that depicts and recognizes information classes or ideas. The model is inferred dependent on the investigation of a lot of preparing Breast Cancer data (i.e., data objects for which the class marks are known). The model is utilized to foresee the class name of items for which the class name is having the breast cancer malady or not having breast cancer ailment that is obscure.

Machine Learning examines how computers can learn (or improve their exhibition) in view of Breast Cancer information. The primary research zone is for computer projects to consequently figure out how to perceive complex examples and settle on clever choices dependent on Breast Cancer data.

Breast Cancer is very important health issue in women needs to have very much need to take care. There has been a lot of research on cancer diagnosis by using machine learning techniques.

In this literature survey, the first paper is the ‘Prediction of Breast Cancer Disease using Machine Learning Algorithms’ by Muktevi Srivenkatesh [1] which proposes a prediction model to predict whether a people have a breast cancer disease or not, by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Random forest, Naive Bayes classifier and logistic regression on the dataset taken in a region to present an accurate model of predicting breast cancer disease.

Hiba Asri, Hajar Mousannif, Hassan Al Moatassime,Thomas Noel [2] has discussed breast cancer analysis with different machine learning algorithms such as Support Vector Machine (SVM), Decision Tree, Naive Bayes (NB) and k-Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted and their performance was compared. Their experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. Their study is also mainly referred and incorporated in this project and the highest accuracy is shown by SVM (96.4%).

The study by Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi [3] is based on genetic programming and machine learning algorithms. Their aim was to construct a system which can accurately differentiate between benign and malignant breast tumors while optimizing the learning algorithm. They have applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. The have presented study and proves that genetic programming can automatically find the best model by combining feature pre-processing methods and classifier algorithms.

Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O Chappuis, Ivo D. Dinov & Maria C. Katapodi [4], their study was to compare the discriminatory accuracy of ML-based estimates against a pair of established methods—the Breast Cancer Risk Assessment Tool (BCRAT) and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) models. Their results showed that their Predictive accuracy reached 90.17% using ML-adaptive boosting and 89.32% using ML-Markov chain Monte Carlo generalized linear mixed model versus 59.31% with BOADICEA for the Swiss clinic-based sample.

Ch. Shravya, K. Pravalika, Shaik Subhani [5] has studied on the implementation of models using Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbor (KNN) is done on the breast cancer dataset taken from the UCI repository. With respect to their results of accuracy, precision, sensitivity, specificity and False Positive Rate the efficiency of each algorithm was measured and compared. Their experimental results have shown that SVM is the best for predictive analysis with an accuracy of 92.7%. They infer that SVM is the well-suited algorithm for prediction and overall KNN presented well after SVM.

Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi [6] their aim is to classify whether the breast cancer is benign or malignant and predict the recurrence and non-recurrence of malignant cases after a certain period. They have used machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes. These techniques are coded in MATLAB using UCI machine learning depository. They have compared the accuracies of different techniques and observed the results. They have found SVM most suited for predictive analysis and KNN performed best for the overall methodology.

This section summarizes some of the scholarly and research works in the field of Machine Learning to correctly predict breast cancer tumors and give the best prediction model with maximum accuracy.

Chapter 3

ANALYSIS

The main purpose of this project is to train machine learning models to predict whether a breast cancer cell is Benign or Malignant. Data will be transformed and its dimension reduced to reveal patterns in the dataset and create a more robust analysis. The optimal model will be selected following the resulting accuracy, sensitivity, and f1 score, amongst other factors. All necessary features are extracted from the Wisconsin dataset and it would be helpful to determine whether a given sample appears to be Benign ("B") or Malignant ("M"). The machine learning models that are applicated in this project report try to create a classifier that provides a high accuracy level combined with a low rate of false-negatives (high sensitivity).

3.1 Scope

The key part for predicting whether a breast cancer cell is Benign or Malignant by working on the Wisconsin dataset and experiment different approaches and building models for various Machine Learning algorithms. To do so, data must be obtained and a dataset over which the experiments will be carried out, based on the domain of interest must be procured, and cleaned.

3.2 Motivation

Breast Cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. The main motivation for the selection of this project was to build a model with the highest accuracy of prediction of the breast cancer cell type – benign or malignant as this would help healthcare facilities treat breast cancer while it is still in the early stage.

3.3 Software Requirement Specifications

A Software Requirements Specification (SRS) is a description of a software system to be developed. It lays out functional and non-functional requirements and may include a set of use cases that describe user interactions that the software must provide.

3.3.1 Overall Description

The proposed system has been built to support the users to have better access and it is also easy to use. The application is open to all users. Machine Learning techniques such as Supervised Learning is best used in predictor models such as the proposed system. The algorithms are compared against each other by testing over the same dataset and the resulting model with the highest accuracy is said to be the best model for the prediction of the type of breast cancer cells.

3.3.2 Functionality

There is only one user-access level:

- End-user

3.3.2.1 Security Requirements

It is of utmost importance to ensure that there is protection against unauthorized access.

3.3.2.2 Performance Requirements

The PCs used must be at least be INTEL CORE i3 machines so that they can give optimum performance of the system. In addition to these requirements, the system should also embrace the following requirements:

- **Reliability:** The system should have little or no downtime.
- **Ease of Use:** The general and administrative views should be easy to use and intuitive.

3.3.2.3 Hardware Requirements

The Hardware requirements are very minimal and the program can be run on most of the machines.

Processor	:	Pentium IV processor
Processor Speed	:	2.4 GHz
RAM	:	1 GB
Monitor Resolution	:	1024*768 or 1336*768 or 1280*102

3.3.2.4 Software Requirements

1. Operating System: Windows 10 / MacOS
2. Web browser: Google Chrome / Safari
3. Anaconda Navigator
4. Jupyter Notebook
5. Python 3
6. Libraries: numpy, pandas, sklearn, seaborn, matplotlib

3.3.3 Dataset

For this project, the Wisconsin breast cancer dataset was used which is available on Kaggle. This dataset is an example of a classification problem in Supervised Machine Learning.

There are 569 rows and 33 columns/features in the dataset, in which 357 are of type Benign ("B") and 212 are Malignant ("M").

The dataset is checked for missing/NA data values and necessary steps are taken. In the data there is a high correlation between certain attributes such as radius mean and worst mean; indicating that the malignant tumors have a high radius.

Exploring further, dependencies among several features are found, thus introducing the concept of multicollinearity. Multicollinearity can be eliminated by removing the dependent features altogether.

Chapter 4

DETAILED SYSTEM DESIGN

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

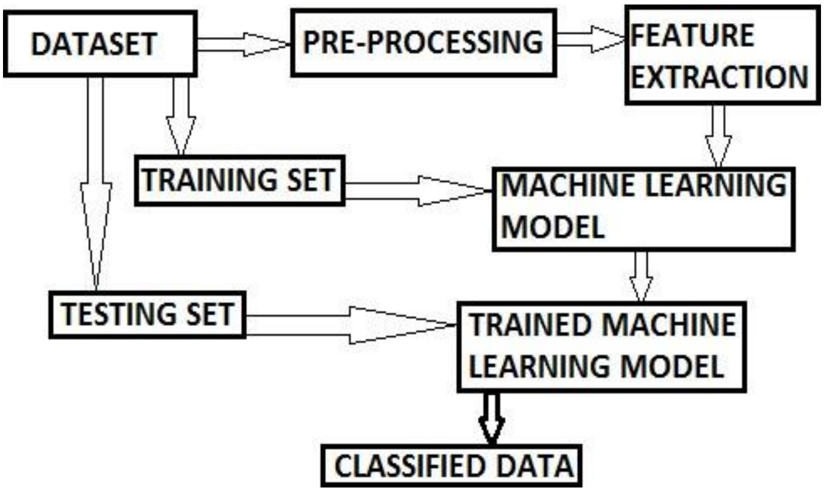


Figure 4.1 Detailed design

The Figure 4.1 shows how the model learns from train data and predicts classes for the test data. The dataset is pre-processed followed by extraction of features used to build the learning model. The part of dataset used to test the model is fed into the trained model. This model further predicts and classifies the data.

4.1 Python Notebook

The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components:

- **A web application:** A browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
- **Notebook documents:** A representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

4.2 Supervised Machine Learning

The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output Y = f(X). The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Techniques of Supervised Machine Learning algorithms include linear and logistic regression, Decision Trees and Support Vector Machines. Supervised learning requires that the data used to train the algorithm is already labelled with correct answers. For example, a classification algorithm will learn to identify animals after being trained on a dataset of images that are properly labelled with the species of the animal and some identifying characteristics.

Supervised learning problems can be further grouped into Regression and Classification problems. Both problems aim to build a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification.

4.3 Algorithms

There are many methods and algorithms to implement a breast cancer prediction system. Supervised Learning algorithms are mainly used and the accuracy of each algorithm is noted down after testing the model.

4.3.1 Classification Algorithms

The classification step usually involves a statistical model like Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, K Nearest Neighbors:

Logistic Regression

- It is one of the most popular classification algorithm in machine learning.
- The logistic regression model describes relationship between predictors that can be continuous, binary, and categorical. Dependent variable can be binary.
- Based on some predictors it is predicted whether something will happen or not. Estimate the probability of belonging to each category for a given set of predictors.

Decision Tree Classifier

Decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, the population or sample is split into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

Decision trees use multiple algorithms to decide to split a node in two or more sub- nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, the purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

Figure 4.2 below represents a decision tree with the root, decision and terminal nodes.

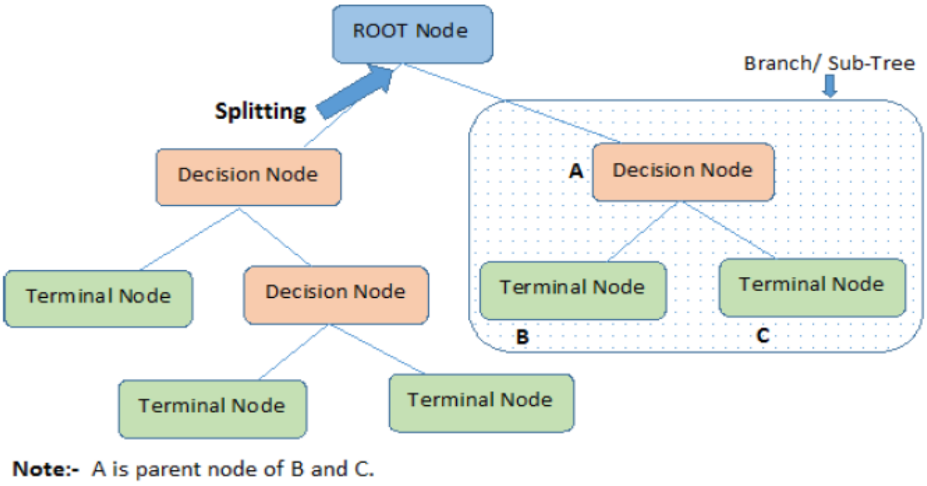


Figure 4.2 Decision Tree

Random Forest Classifier

Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of overfitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provides a good estimate of the generalization error and to be resistant to overfitting.

Support Vector Machine Classifier

- SVM is a one of the machine learning algorithms for regression, classification. It is a supervised learning algorithm that analyses data used for classification and regression.
- SVM modelling involves two steps, firstly to train a data set and to obtain a model & then, to use this model to predict information of a testing data set.
- SVM is a discriminative classifier formally defined by a separating hyperplane where the model represents the training data points as points in space and mapping is done so the points which are of different classes are divided by a gap that is as wide as possible.
- Mapping is done in to the same space for new data points and then predicted on which side of the gap they fall.
- In SVM algorithm, plotting is done as each data item is taken as a point in n-dimensional space where n is number of features with the value of each feature being the value of a particular coordinate. Then, classification is performed by locating the hyper-plane that separates the two classes very well.

  Figure 4.3 below is a Support Vector Machine graph that splits the two classes with an optimal hyperplane.

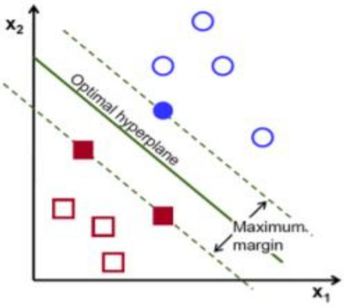


Figure 4.3 Support Vector Machine classification graph

K Nearest Neighbors

- The concept of nearest neighbor analysis has been used in several anomaly detection techniques.
- One of the best classifier algorithms that have been used in the credit card fraud detection is k- nearest neighbor algorithm that is a supervised learning algorithm where the result of new instance query is classified based on majority of K-Nearest Neighbor category.
- The performance of KNN algorithm is influenced by three main factors:
  i. The distance metric used to locate the nearest neighbors.
  ii. The distance rule used to derive a classification from k nearest neighbors.
  iii. The number of neighbors used to classify the new sample.

4.4 Python Libraries

4.4.1 Scikit-Learn

The Scikit-learn project started as scikits.learn, a Google Summer Code project by David Cournapeau. It is a powerful library that provides many machine learning classification algorithms, efficient tools for data mining and data analysis. Below are various functions that can be performed using this library:

- Classification: Identifying the category to which a particular object belongs.
- Regression: Predicting a continuous-valued attribute associated with an object.
- Clustering: Automatic grouping of similar objects into sets.
- Dimension Reduction: Reducing the number of random variables under consideration.
- Model selection: Comparing, validating and choosing parameters and models.
- Pre-processing: Feature extraction and normalization in order to transform input data for use with machine learning algorithm. In order to work with scikit-learn, NumPy is installed on the system.

4.4.2 NumPy

NumPy is the fundamental package for scientific computing with Python. It provides a high performance multidimensional array object, and tools for working with these arrays. It contains among other things:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

4.4.3 Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three clause BSD license.[2] The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Pandas is mainly used for machine learning in form of data frames. Pandas allow importing data of various file formats such as csv, excel etc. Pandas allows various data manipulation operations such as group by, join, merge, melt, concatenation as well as data cleaning features such as filling, replacing or imputing null values. Functionalities of pandas are:

- Data Frame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and sub setting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower dimensional data structure.

4.5 Setting Up Environment for Breast Cancer Prediction

The following components are required to be downloaded and installed properly:

- Download and install Python 2.6 or above in a desired location.
- Download and install NumPy and pandas libraries.

Chapter 5

IMPLEMENTATION

Implementation is the process of defining how the system should be built, ensuring that it is operational and meets quality standards. It is a systematic and structured approach for effectively integrating a software-based service or component into the requirements of end users.

5.1 Breast Cancer Predictor

A breast cancer predictor can be applied to predict the breast cancer cell type based on certain features. The predictor aims at classifying the cells as benign or malignant. It can also be enhanced to correctly predict and provide a diagnostic approach to treat breast cancer (malignant). The model takes an input of data values with features that describe the type of cancer cell and predicts based on the test data that is provided. The model is trained using Kaggle dataset from Wisconsin.

5.2 Python

Python is an interpreted high-level programming language for general-purpose programming and has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution. Rather than having all its functionality built into its core, Python was designed to be highly extensible. Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation.

5.3 Programming Coding Guidelines

Following Code guidelines are important to programmers for several reasons:

- 80% of the lifetime cost of a piece of software goes to maintenance.
- Hardly any software is maintained for its whole life by the original author.

- Code conventions improve the readability of the software, allowing engineers to understand new code more quickly and thoroughly.

If you ship your source code as a product, you need to make sure it is as well packaged and clean as any other product you create.

5.4 Methodology of the project

It is ensured that a systematic and structured way of designing the project in a very transparent manner for achieving the objective goals, the representation of the methodology followed here is shown in Figure 5.1 below.

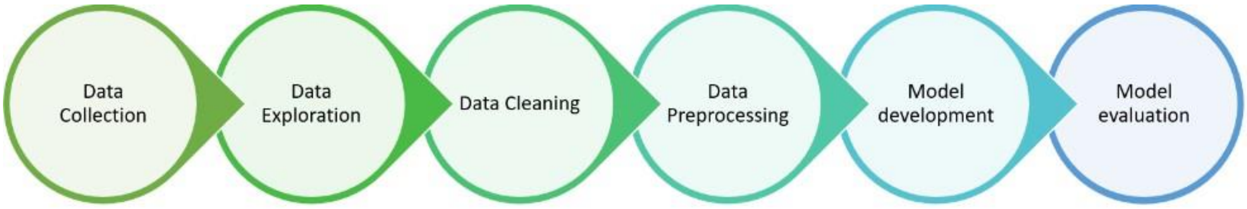


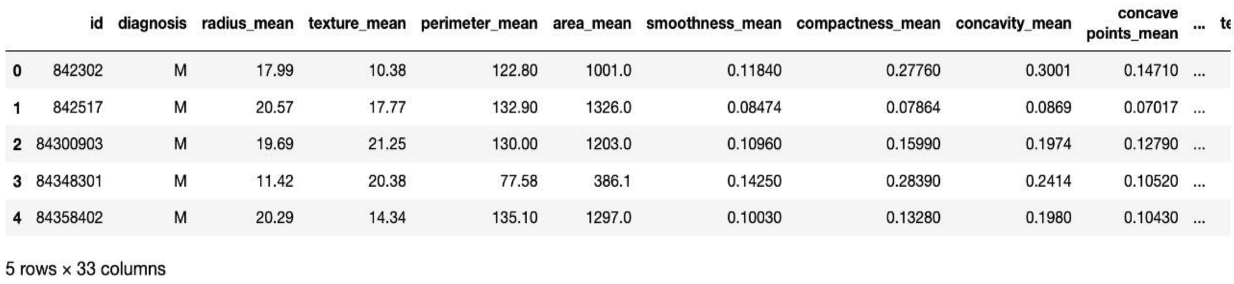
Figure 5.1 Project Methodology

Every step in Figure 5.1 is rigorously followed and consider it as a standard procedure for building the machine learning projects, this is explained in brief in future topics.

5.4.1 Data Collection and Exploration

The data requirements of the proposed model can be divided into two:

- **Training data:** The training data has been taken from Kaggle. It is a .csv file containing 569 rows of which the first 5 rows are shown in Figure 5.2. It also shows the training data consisting of 33 columns- id, diagnosis (B or M), radius_mean, texture_mean, perimeter_mean, area_mean, and 26 other columns/features.



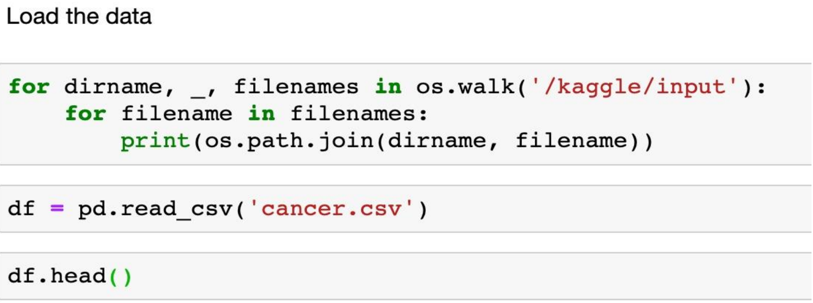
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	te
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 33 columns

Figure 5.2 Training Data

- **Test data:** The test data is a part of the Wisconsin dataset that is taken from Kaggle. The split percentage of test data from the dataset is 30% (70% is taken as training data).

Here, the data set (csv file) has been imported into the working python area (Jupyter notebook) for performing various techniques. The code snippet for the importing dataset into the notebook is given in the Figure 5.3 below. Data exploring refers to view the dataset in an organized way.



Load the data

```
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
df = pd.read_csv('cancer.csv')
```

```
df.head()
```

Figure 5.3 Data Exploration

5.4.2 Data Cleaning

Since the model deals with reading and classifying by predicting the type of breast cancer cell, it should be made easy for it to do so. Data cleaning involves steps to clean the data bringing it to the format required by the model. There are several python libraries the make pre-processing possible. Some of the common cleaning methods are:

- **Removal of unnecessary features:** The features that do not play a role in the prediction of the breast cancer system are removed (for example, id).
- **Removal of NaN data values:** Features which have only NaN data values in the dataset are not needed and hence are removed.
- **Mapping of the two cell types:** In order to make sure that the system works properly, the data objects ('B' and 'M' values in diagnosis) are mapped to 0 and 1 ('B': 0 and 'M': 1).
- **Check for missing/null values:** Cleaning of dataset is crucial because it may contain null values. Thus, it becomes necessary to check for missing values and replace them.
- **Check for multicollinearity and remove dependent features:** There may be some features that have high correlation with another feature. These features may be dependent on the other and does not have any significance in the dataset. So removing these dependent features is a part of cleaning the data in order to improve efficiency.

5.4.3 Data Pre-processing and Model development

This is the most important part in the machine learning workflow. Since the algorithm is totally dependent on how the data is fed into it, feature engineering which is an integrated step in data pre-processing should be given top most priority for every machine learning project.

Some of the advantages of the pre-processing of the data is that it reduces Overfitting i.e. less redundant data will be possible which means less opportunity to make decisions based on noise and it improves accuracy which means less misleading data and thus the modelling accuracy improves. Training Time is reduced i.e. fewer data points reduce algorithm complexity and algorithms train faster.

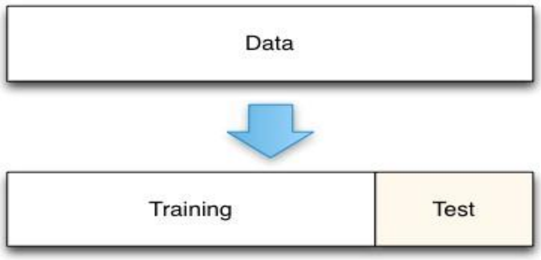
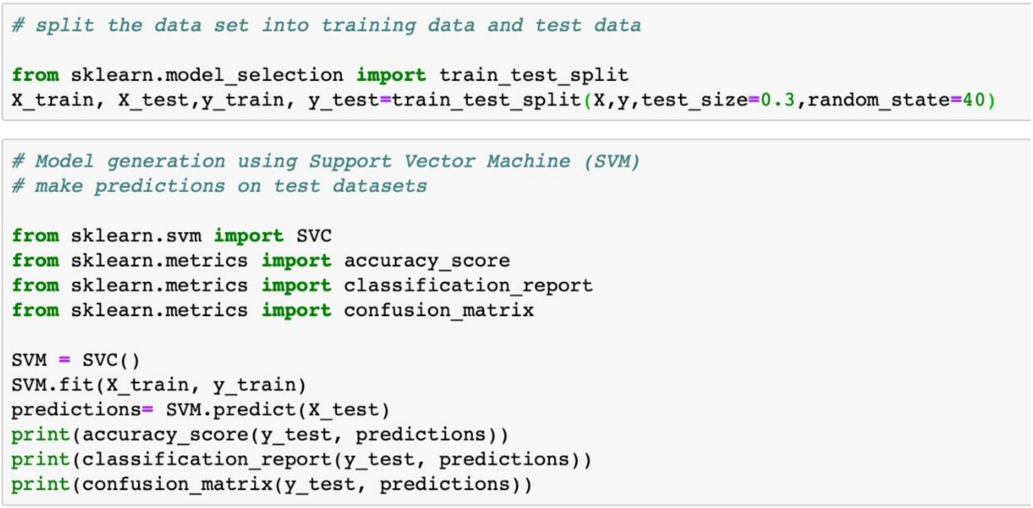


Figure 5.4 Data Pre-processing

Figure 5.4 shows how the data is split into training and test data (70% training data and 30% test data).

Model development plays an important role in making the project in a comprehensive way by splitting the variables as outcome variable and predicting the project creates a separate function which tends to be called when a technique is used. A confusion matrix which provides the information regarding the attributes is displayed.

In the below snippet, Figure 5.5 a function is created which takes the parameters as model name, the data and predicting values and outcome values, it uses fitting mechanism for organizing the data and finally it prints the predictions calculated. The data being used is usually split into outcome variable and predictor variables. The outcome set contains a known output and the model learns on this data in order to be generalized to other data later on.



```
# split the data set into training data and test data

from sklearn.model_selection import train_test_split
X_train, X_test,y_train, y_test=train_test_split(X,y,test_size=0.3,random_state=40)
```

```
# Model generation using Support Vector Machine (SVM)
# make predictions on test datasets

from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix

SVM = SVC()
SVM.fit(X_train, y_train)
predictions= SVM.predict(X_test)
print(accuracy_score(y_test, predictions))
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions))
```

Figure 5.5 Model Development

5.4.4 Model Evaluation

The accuracy of the model is evaluated to see how well the type of breast cancer cell is predicted. This information can later be used to compare with other techniques and choose the best model for deployment.

Support Vector Machine Algorithm

Support vector machine algorithm aims to find a hyperplane in an N-dimensional space (N- the number of features) that distinctly classifies the data points. To separate the two classes of data points there are many possible hyperplanes that could be chosen. A plane that has the maximum margin is taken i.e. the maximum distance between the data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified correctly. SVM shows the highest accuracy among all the classifier algorithms in this project.

Chapter 6

TESTING

6.1 Unit Testing

Unit testing is the level of software testing where the individual units/components of a software are tested. The purpose is to validate each unit of the software performs as desired. Here the validation concepts are used to check whether the program is taking the inputs in the correct format or not.

Table 6.1 shows the unit testing for dropping a specified column by using drop function.

Table 6.1 Unit test case for dropping a column function

Sl. No. of test case	1
Name of test	Unit Testing
Item / Feature being tested	Dataset
Sample Input	Drop command to drop a column in the table
Expected output	Column gets dropped
Actual output	Specified column gets dropped
Remarks	Test succeeded

6.2 Integration Testing

Integration testing is also taken as integration and testing this is the major testing process where the units are combined and tested. Its main objective is to verify whether the major parts of the program is working fine or not. This testing can be done by choosing the options in the program and by giving suitable inputs it is tested. Table 6.2 shows integration testing.

Table 6.2 Integration test case for splitting the data

Sl. No. of test case	1
Name of test	Integration testing
Item / Feature being tested	Train-test split functionality
Sample Input	Input features and ratio and check for the correct splitting of the dataset.
Expected output	Split the dataset according to given ratio.
Actual output	Split the dataset according to given ratio.
Remarks	Test succeeded

6.3 System Testing

System Testing (ST) is a black box testing technique performed to evaluate the complete system the system's compliance against specified requirements. The software or hardware is testing conducted on a complete, integrated system to evaluate the system’s compliance with its specified requirements. Tables 6.3, 6.4 show system test cases for predictions made by the breast cancer predictor model on different learning algorithms (SVM and Random Forest classifier).

Table 6.3 System test case for prediction using SVM

Sl. No. of test case:	1
Name of test:	System testing
Item / Feature being tested:	Final model's accuracy in predicting breast cancer using Support Vector Machine (SVM)
Sample Input:	Data with the cell features
Expected output:	Good accuracy of breast cancer prediction
Actual output:	Accuracy of 96.4% (Highest among all classification algorithms)
Remarks:	Test succeeded

Table 6.4 System test case for prediction using Random Forest

Sl. No. of test case	2
Name of test	System testing
Item / Feature being tested	Final model's accuracy in predicting breast cancer using Random Forest classifier
Sample Input	Data with the cell features
Expected output	Accuracy of breast cancer prediction
Actual output	Accuracy of 92.9%
Remarks	Test succeeded

Chapter 7

RESULTS AND SNAPSHOTS

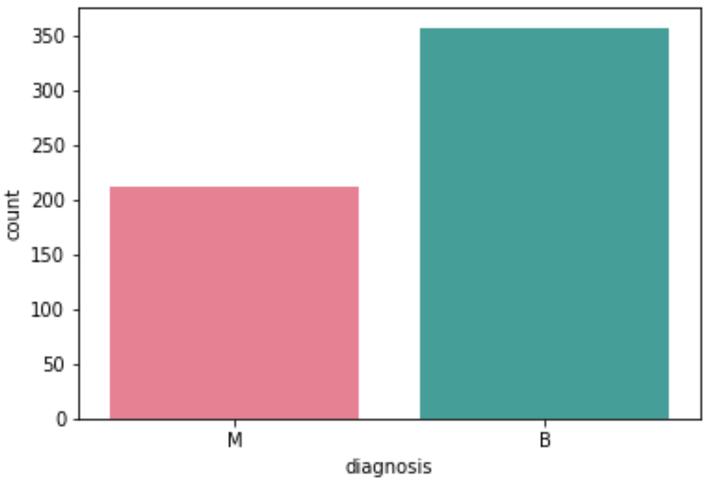


Figure 7.1 Count of Benign and Malignant cases

The above Figure 7.1 is a count plot that shows the number of rows that have ‘M’ (Malignant – 212) and ‘B’ (Benign – 357) in the diagnosis column/feature.

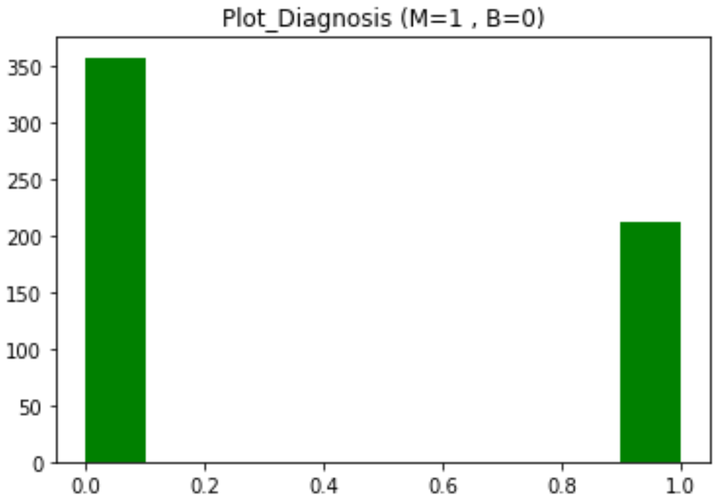


Figure 7.2 Histogram of Diagnosis column

Figure 7.2 displays a histogram with data values ‘M’ and ‘B’ from the diagnosis column that are mapped to 1 and 0, respectively and plotted using a histogram.

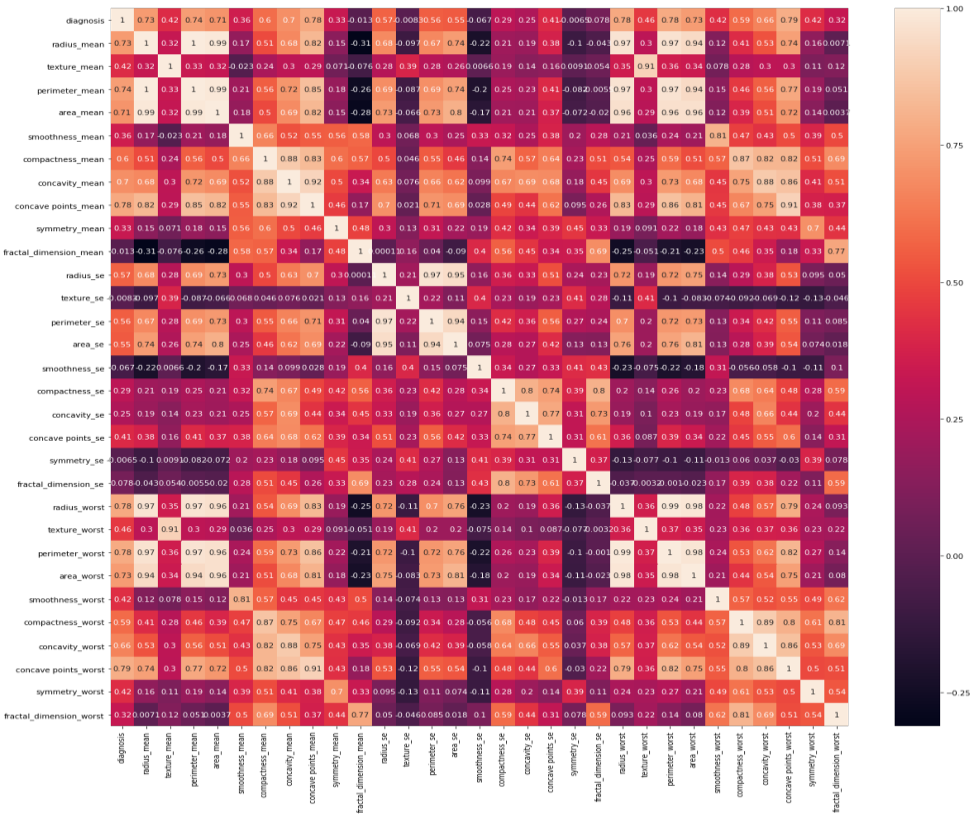


Figure 7.3 Heatmap showing multicollinearity among features

The Figure 7.3 represents a heatmap that shows the high correlation present between the columns and implies that they are dependent on each other, and need to be removed.

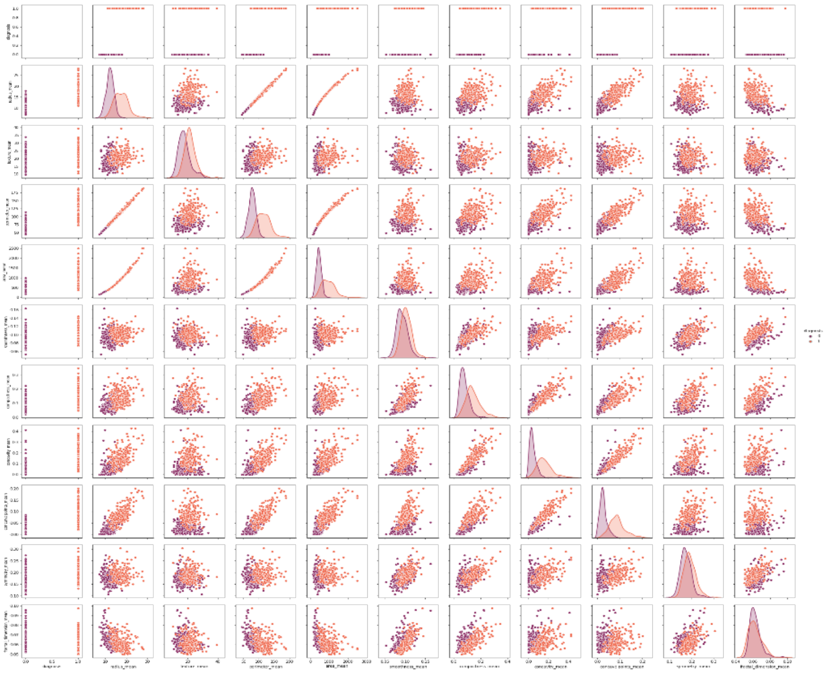


Figure 7.4 Pairplot showing dependency between features

Figure 7.4 displays a pair plot with linear patterns that show how some features are dependent on one another.

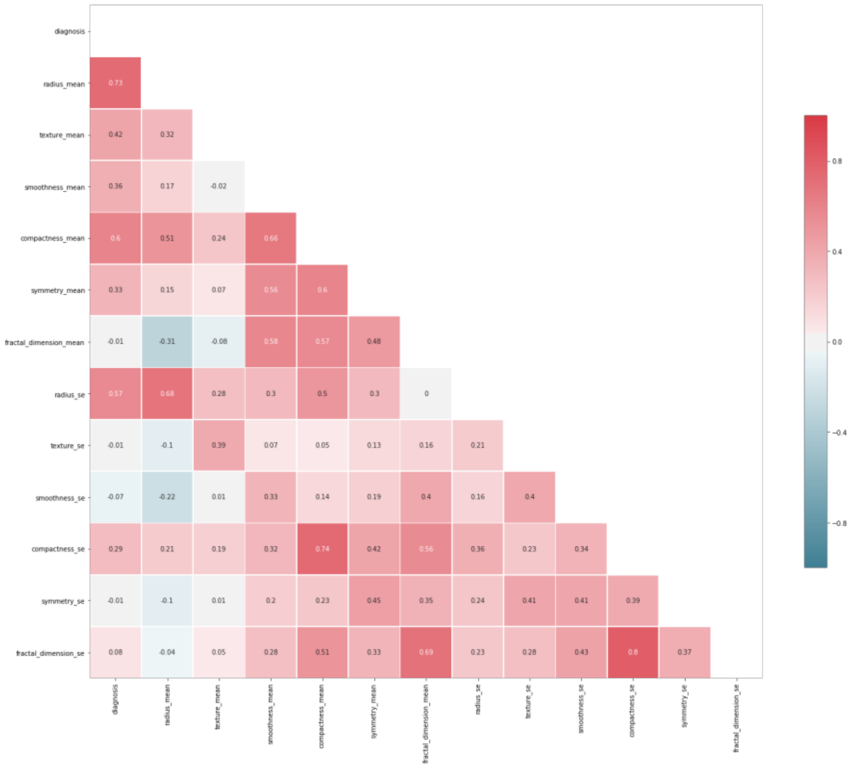
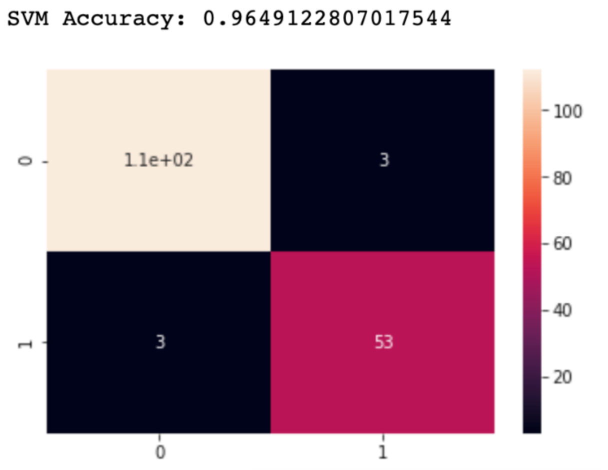


Figure 7.5 Heatmap after removing multicollinearity

Figure 7.5 is a heatmap shows how the features no longer depend on another feature, i.e. the highly-correlated predictors are removed.



```
SVM Accuracy: 0.9649122807017544
```

Figure 7.6 Confusion matrix

Figure 7.6 is a confusion matrix showing the number of actual and false positives, negatives predicted by the SVM model. Based on the difference in true label and predicted label, accuracy of the model is found to be around 96.4% which is the highest accuracy among various classification algorithms.

Chapter 8

CONCLUSION AND FUTURE ENHANCEMENT

In this report, different machine learning algorithms are reviewed for the prediction of breast cancer. The main focus is to find out the most suitable algorithm that can predict the breast cancer tumor cell more effectively and correctly. The main purpose of this project report is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction. The review of major machine learning techniques and ensemble techniques has been provided and these techniques deeply elaborate algorithms that are being used for the predictions of breast cancer.

The accuracy obtained by Support Vector Machine (SVM) is the highest (96.4%) among all the classification algorithms i.e. Logistic Regression (95.9%), Decision Tree Classifier (89.4%), Random Forest Classifier (92.9%), k-Nearest Neighbors (96.4% , but the number of incorrectly rejected cases are greater in KNN) and Naïve Bayes (92.9%).

In the future work, there are still some issues that needed to be solved. Researchers can solve the issue of limited available dataset by using some data augmentation techniques. The issue of inequality of positive and negative data should be considered by researchers as it can lead to biasness towards positive or negative prediction.

The future enhancement of this project can be done by extending the project to detect breast cancer at early stages as to provide proper diagnosis, build other models using deep learning techniques, etc., and optimize the learning algorithms.

Hence, to conclude, it was a great experience, precisely a learning experience to build this project and to understand the various requirements of a successful project consisting of many concepts that are present in it.

REFERENCES

[1] Dr. Muktevi Srivenkatesh, “Prediction of Breast Cancer Disease using Machine Learning Algorithms”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-9 Issue-4, February 2020.

[2] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (2016) 1064 – 1069.

[3] Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi, “Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms”, Journal of Health Care Engineering, Volume 2019.

[4] Chang Ming, Valeria Viassolo, Nicole Probst-Hensch, Pierre O Chappuis, Ivo D. Dinov & Maria C. Katapodi, Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models, Breast Cancer Research volume 21, Article number: 75 (2019).

[5] Ch. Shravya, K. Pravalika, Shaik Subhani, “Prediction of Breast Cancer Using Supervised Machine Learning Techniques”, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-6, April 2019.

[6] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, “Breast cancer diagnosis and recurrence prediction using machine learning techniques”, International Journal of Research in Engineering and Technology Volume-4, Issue-4, April 2015.

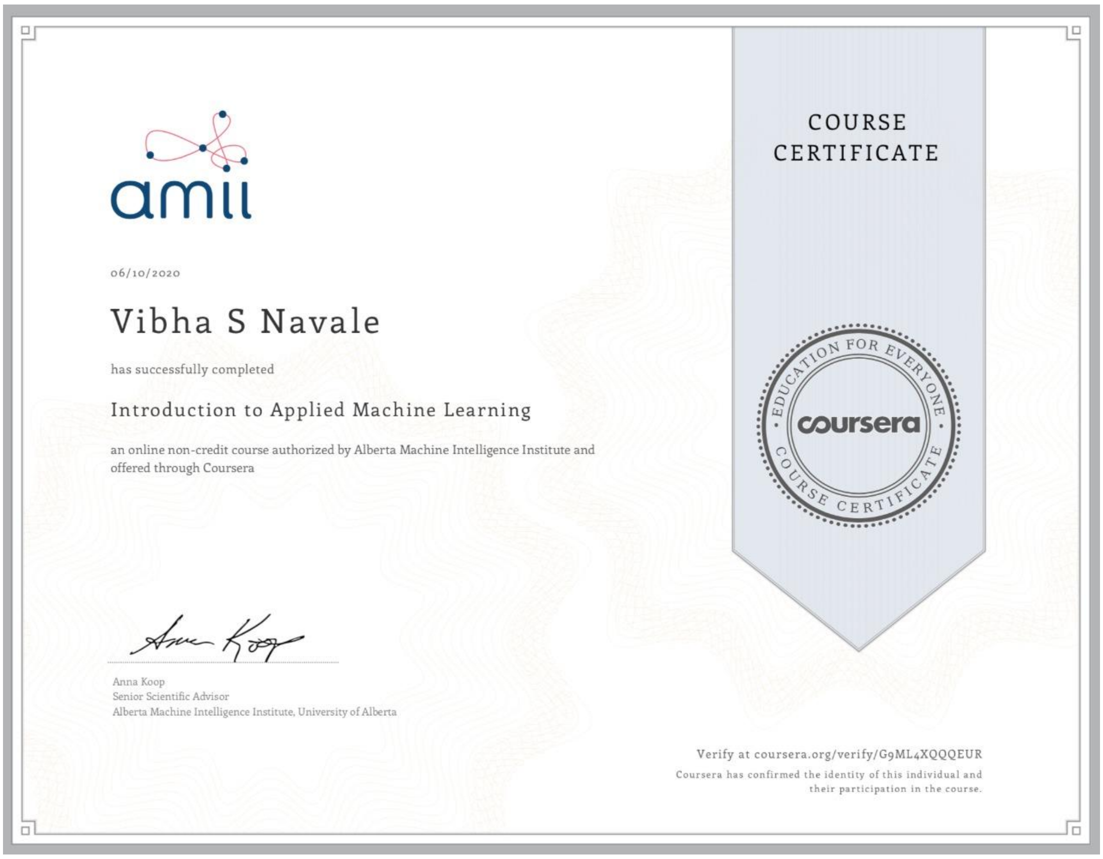
[7] Ratula Ray, Azian Azamimi Abdullah, Debasish Kumar Mallick, Satya Ranjan Dash, “Classification of Benign and Malignant Breast Cancer using Supervised Machine Learning Algorithms Based on Image and Numeric Datasets”, International Conference on Biomedical Engineering (ICoBE), 2019.

[8] Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018.

[9] Hazra, A., Mandal, S.K., Gupta, A.: Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and Ensemble Algorithms. Int. J. Comput. Appl. **145**, 0975–8887 (2016).

[10] UCI Machine Learning Breast Cancer Wisconsin Dataset on Kaggle.

COURSERA CERTIFICATES



amii

COURSE CERTIFICATE

06/10/2020

Vibha S Navale

has successfully completed

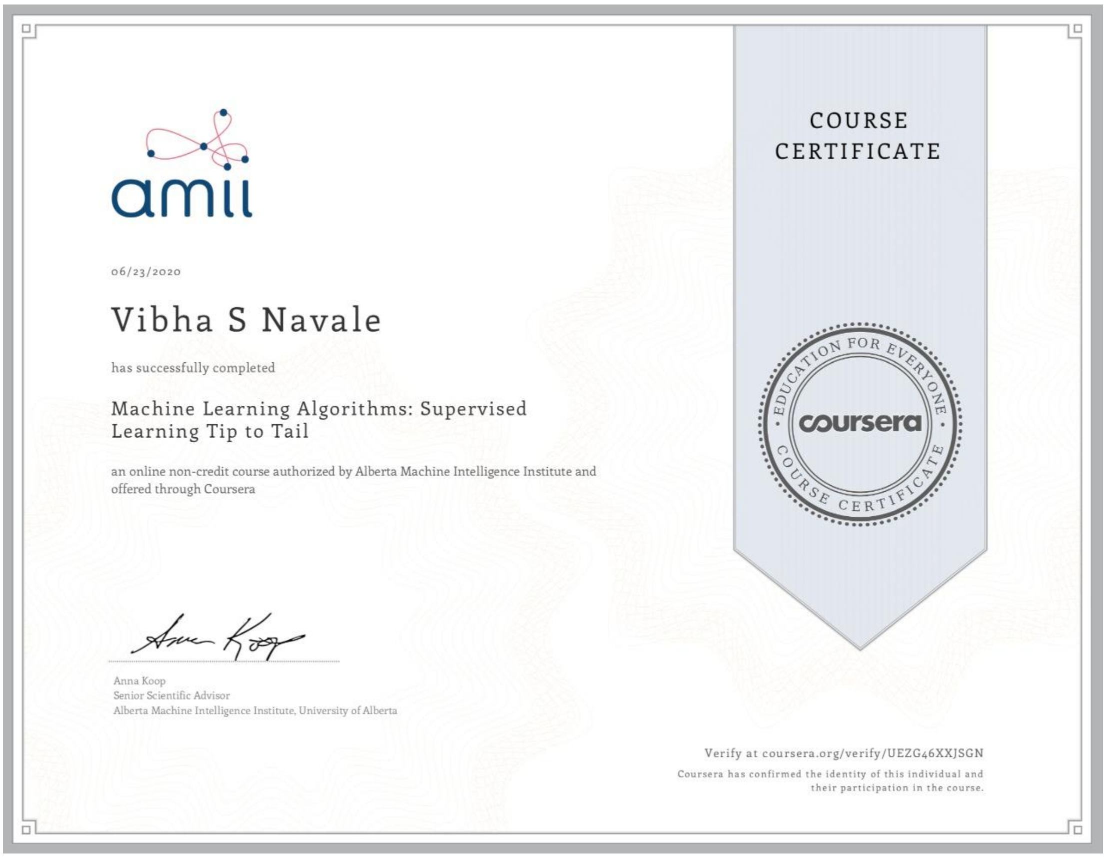
Introduction to Applied Machine Learning

an online non-credit course authorized by Alberta Machine Intelligence Institute and offered through Coursera

EDUCATION FOR EVERYONE coursera COURSE CERTIFICATE

Anna Koop
Senior Scientific Advisor
Alberta Machine Intelligence Institute, University of Alberta

Verify at coursera.org/verify/G9ML4XQQQEUR
Coursera has confirmed the identity of this individual and their participation in the course.



amii

COURSE CERTIFICATE

06/23/2020

Vibha S Navale

has successfully completed

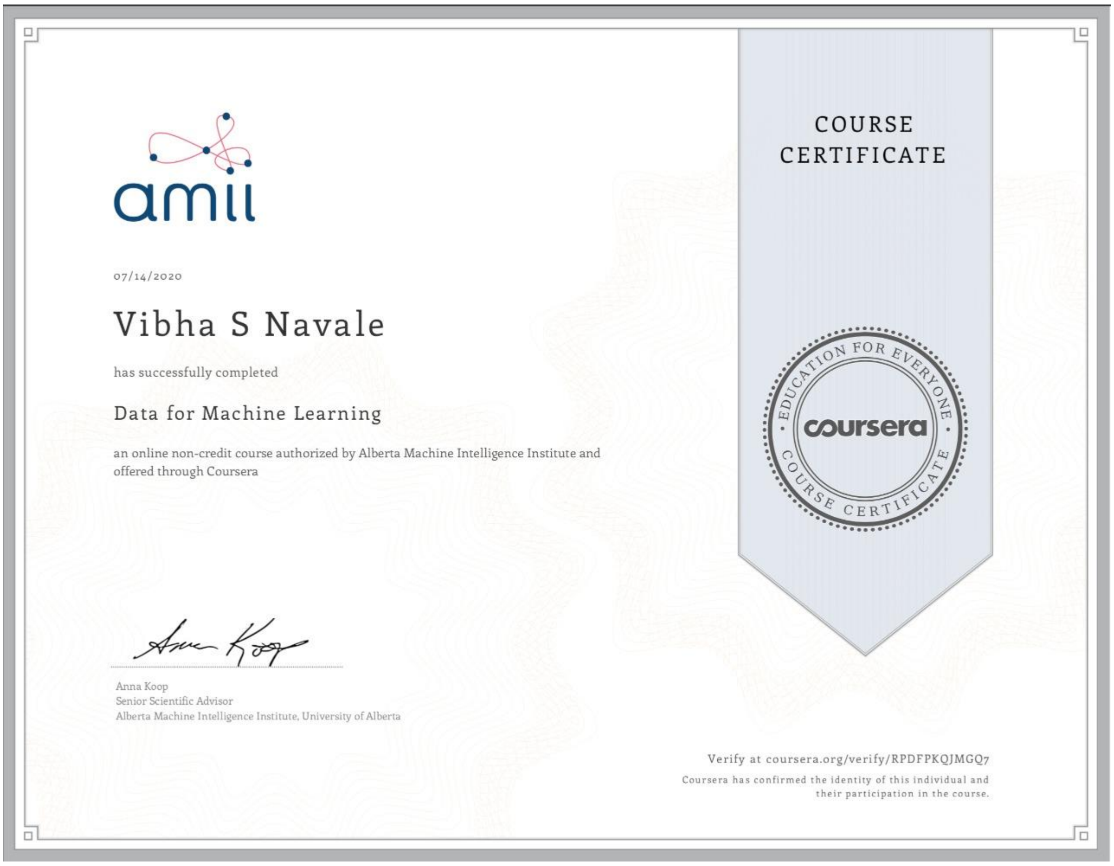
Machine Learning Algorithms: Supervised Learning Tip to Tail

an online non-credit course authorized by Alberta Machine Intelligence Institute and offered through Coursera

EDUCATION FOR EVERYONE coursera COURSE CERTIFICATE

Anna Koop
Senior Scientific Advisor
Alberta Machine Intelligence Institute, University of Alberta

Verify at coursera.org/verify/UEZG46XXJSGN
Coursera has confirmed the identity of this individual and their participation in the course.



amii

COURSE CERTIFICATE

07/14/2020

Vibha S Navale

has successfully completed

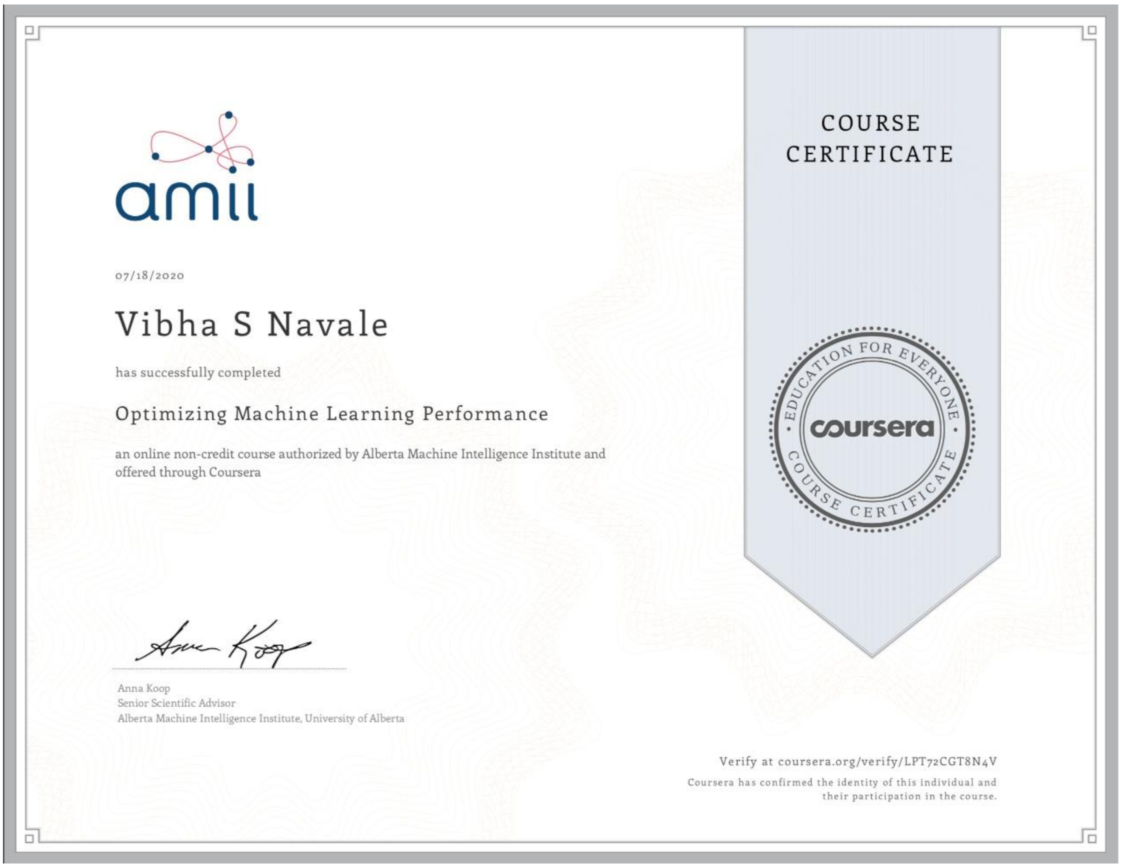
Data for Machine Learning

an online non-credit course authorized by Alberta Machine Intelligence Institute and offered through Coursera

EDUCATION FOR EVERYONE coursera COURSE CERTIFICATE

Anna Koop
Senior Scientific Advisor
Alberta Machine Intelligence Institute, University of Alberta

Verify at coursera.org/verify/RPDFPKQJMGQ7
Coursera has confirmed the identity of this individual and their participation in the course.



amii

COURSE CERTIFICATE

07/18/2020

Vibha S Navale

has successfully completed

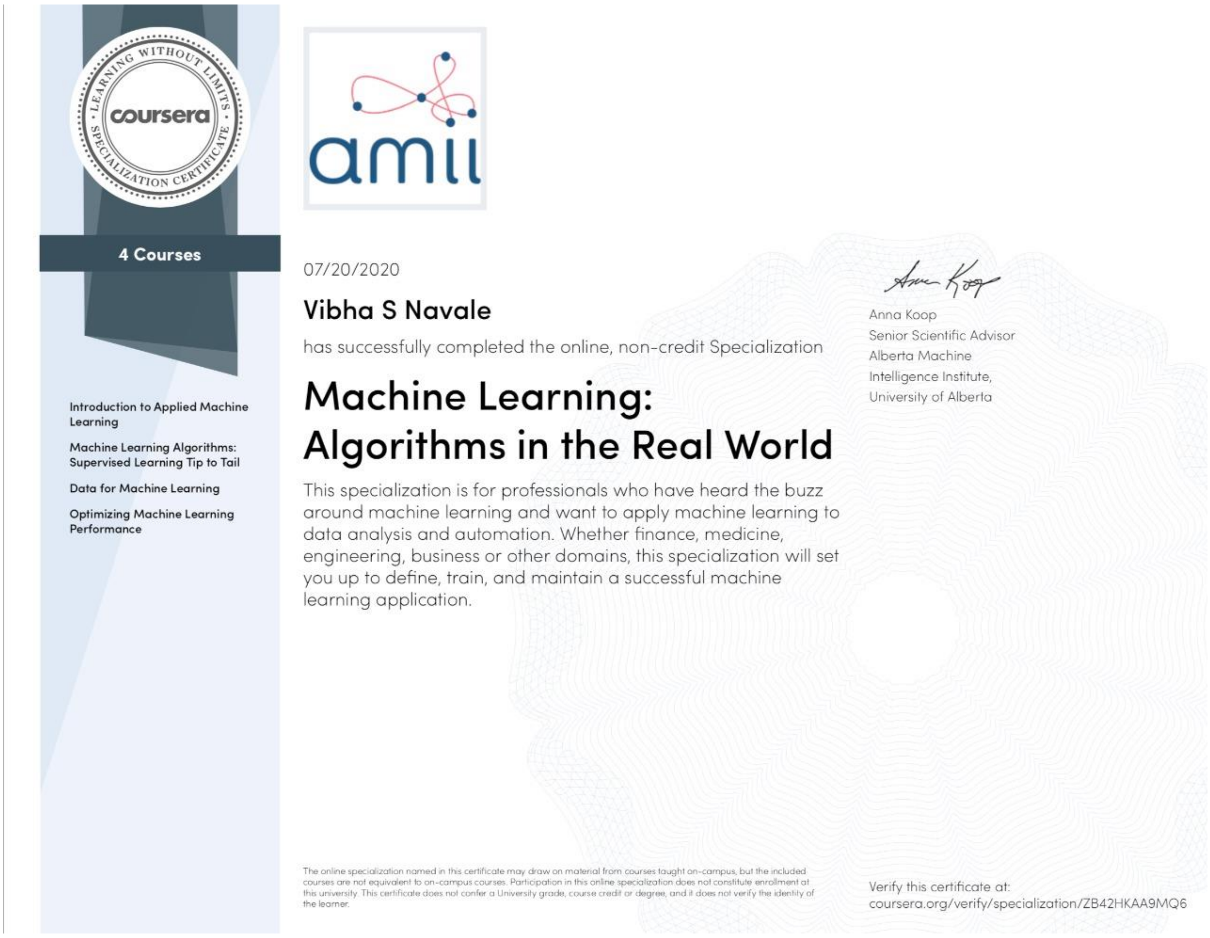
Optimizing Machine Learning Performance

an online non-credit course authorized by Alberta Machine Intelligence Institute and offered through Coursera

EDUCATION FOR EVERYONE coursera COURSE CERTIFICATE

Anna Koop
Senior Scientific Advisor
Alberta Machine Intelligence Institute, University of Alberta

Verify at coursera.org/verify/LPT72CGT8N4V
Coursera has confirmed the identity of this individual and their participation in the course.



LEARNING WITHOUT LIMITS

coursera

SPECIALIZATION CERTIFICATE

amii

4 Courses

Introduction to Applied Machine Learning

Machine Learning Algorithms: Supervised Learning Tip to Tail

Data for Machine Learning

Optimizing Machine Learning Performance

07/20/2020

Vibha S Navale

has successfully completed the online, non-credit Specialization

Machine Learning: Algorithms in the Real World

This specialization is for professionals who have heard the buzz around machine learning and want to apply machine learning to data analysis and automation. Whether finance, medicine, engineering, business or other domains, this specialization will set you up to define, train, and maintain a successful machine learning application.

Anna Koop

Anna Koop
Senior Scientific Advisor
Alberta Machine
Intelligence Institute,
University of Alberta

The online specialization named in this certificate may draw on material from courses taught on-campus, but the included courses are not equivalent to on-campus courses. Participation in this online specialization does not constitute enrollment at this university. This certificate does not confer a University grade, course credit or degree, and it does not verify the identity of the learner.

Verify this certificate at:
coursera.org/verify/specialization/ZB42HKAA9MQ6