**Instructions**: Please complete and submit your work to the appropriate folder in LumiNUS. You may work in study groups, but each student must be responsible for their own submission.

Please submit all the following documents as a single zip file named StudentID-Name-H1.zip:

1. Completed Word file named as StudentID-Name-H1.docx (with all results)
2. Print preview of ipynb file named as StudentID-Name-H1.pdf (with all results)
3. Working ipynb file named as StudentID-Name-H1.ipynb

1. A machine learning application scenario can be seen from four perspectives: (i) what is the technical problem to be solved, (ii) the data requirements, meaning can you get the data needed, (iii) security and privacy considerations, meaning what happens if data is leaked and (iv) the value proposition that machine learning brings to the table.

Choose one industry from the list below (or be creative and come up with your own) and give one example of how supervised machine learning can be applied. Your examples should follow the 4 perspectives outlined above.

Give a different example from those given in lecture. You may consult the Internet, but you must think things through yourself.

|  |  |  |
| --- | --- | --- |
| Retail | Fashion | Industry 4.0 |
| Banking | Education | Social Media |
| Healthcare | Communication Networks | Smart Home |

**ANSWER 1:**

* *Industry Chosen:*

**Healthcare**

* *Technical Problem To Be Solved:*

**Bacteria Identification & Classification For Clinical Research**

* *Working Methodology:*
* **Supervised Machine Learning Algorithms** are used in the case to identify different bacteria species with a certain % of accuracy.
* Labelled data from patients are fed into supervised ML models and outputs which identify certain species of bacteria are identified using a classification technique.
* Labelled data can include the following in my opinion:
  + Bacteria images
  + Patient disease data
  + Bacteria characteristics: size, colour and photo location
* Results from labelled input data could be a classification model where bacteria could be classified into 3 classes: Present, Absent And Unknown for example based on the labelled input dataset on bacteria or categorize them into different classes based on type of disease they cause for instance.
* Supervised ML algorithms can efficiently classify labelled on the basis of feature extraction techniques which could be broken down in a few simple steps:
  + First phase executes the acquisition of bacterial images
  + Second phase performs data-pre-processing to remove unwanted noise, blur etc. through different operations to make the image data more clear and technically suitable for processing.
  + The third phase deals with the segmentation of data.
  + The fourth phase carries out feature extraction and feature selection on digital images
  + Lastly, classification of bacteria images into their respective classes.
* *Data Requirements And Challenges:*
* We can get the data needed primarily from hospital records, especially from patients on which type of bacteria they’ve been infected with so that we can classify the bacteria images to their types using supervised ML for example. We can also get such data from laboratories from clinical research.
* The biggest challenge in data gathering for this particular problem would be the lack of accuracy in medical images analysis, human errors, huge time loss for lab personnel when assessing medical images.
* Another possible challenge for this situation might be that there are limited amounts of bacteria that can be tested using a supervised model as not all of them can be “labelled”, hence, it might be better to have an unsupervised Machine Learning so that the model could be applicable to all sorts of bacteria and would exclude long-term data labeling.
* Poor quality data might be possible which might affect accuracy of the supervised ML model if there are gaps in patient profiles and if hospital records not updated for example.
* *Security And Privacy Considerations:* 
  + Data leak if it occurs would be detrimental as you would be exposing patient health records to the public, such as what disease are they infected with, instead of pure research purposes.
  + There are certain consumer protection and data protection laws in each country which would be detrimental to hospitals if such data leakage occurs.
  + HIPAA and other privacy regulations ensure the security of the patient’s information. Everybody should have a right to keep information about their health private. Nevertheless, a lot of healthcare data leaks are happening every day that result in up to millions of dollars in fine for healthcare providers.
  + The decisions made by the machine learning algorithm completely rely on the data it has been learned on. If the input is unreliable or wrong, the result will be wrong as well (like a garbage in, garbage out method) which might harm the patient if doctors classified the wrong type of bacteria and unable to diagnose the disease that the patient has been affected with.
* *Value Proposition:*

In my opinion, the biggest value proposition that I can think of in this situation is that such Machine Learning applications for Healthcare would be the high precision and accuracy levels in classification and identification of harmful bacteria which would be used for the following purposes which can be high as 99%:

1. Traditional methods of classification are time consuming and frequently prone to errors. To overcome these problems the application of traditional methods creates a wide scope for the scientists to adopt ML approaches in the field of bacteria classification which boosts accuracy for bacteria identification for disease treatment.
2. Ensuring that people are immunized from bacteria identified for disease prevention.
3. Advance clinical trials and developments in the lab for antibiotics.
4. Identify the leading causes of diseases in humans caused by bacteria using data gathered from supervised ML models.
5. Another possible use case what I can think of would be that supervised ML could automate the process of bacteria detection reducing the number of errors made and these complex tasks as bacteria detection is that with ML it is possible to be sure that the same mistake will not be made twice.
6. Redo Problem 1 for the unsupervised learning scenario. Give one example of an application scenario that is different from the examples discussed in the lectures.

**ANSWER 2:**

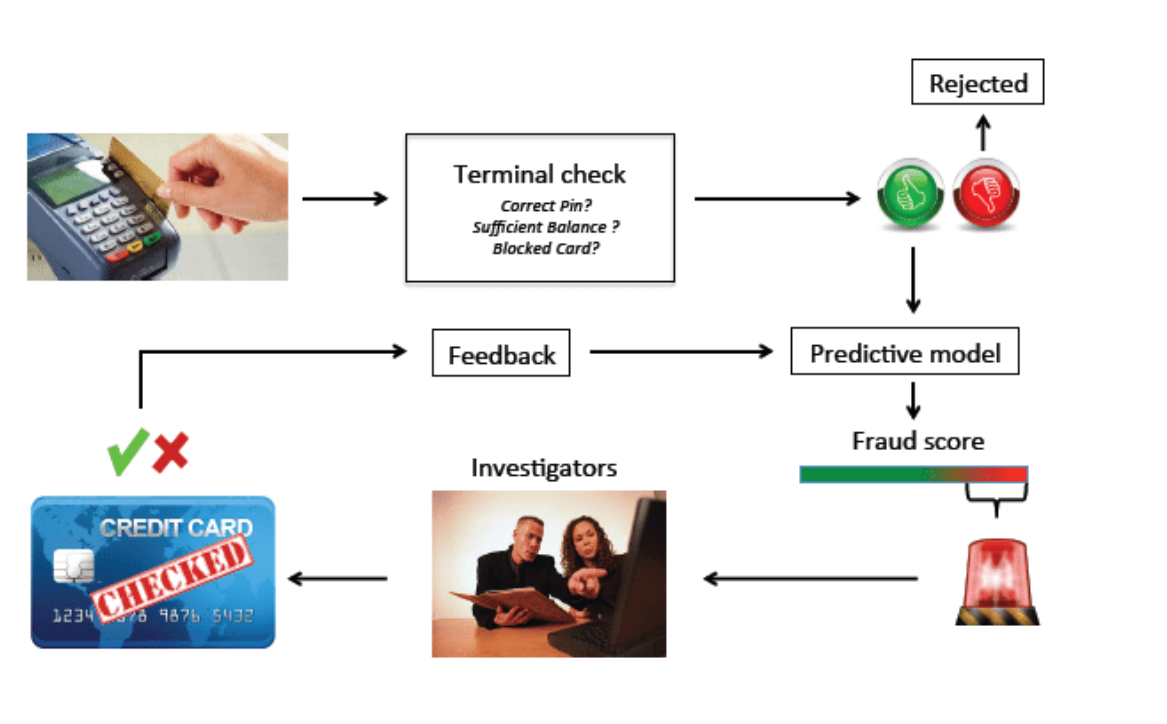
* *Industry Chosen:*

**Banking**

* *Technical Problem To Be Solved:*

**Anomaly Detection In Banking Transactions: Fraud Detection such as Money Laundering issues**

* *Working Methodology:*
* **Unsupervised Machine Learning Algorithms** are used in the case of fraud detection such as money laundering, illegal or suspicious transactions and even for document verification purposes for digital banking purposes.
* The machine learning model employed here works by Anomaly Detection through the various historical transactions such as payments made by bank customers made over a yearly period overall. Algorithm training, validation, and backtesting are based on vast datasets of credit card transaction data. ML-powered classification algorithms can easily label events as fraud versus non-fraud to stop fraudulent transactions in real-time.
* The model used is unsupervised as you are taking data from bank transactions from it’s history (which we do not know whether a fraud has been committed or not), and we are training the data into the ML model and then the data will be classified as fraud or non-fraud transaction, in my opinion.
* Machine learning fraud prevention is among the most effective applications of technology to date. Unsupervised ML models will analyze millions of data points, transaction parameters, and consumer behavior patterns in real-time to detect security threats and potential fraud cases.
* Fraud Detection frameworks are designed to reduce these kinds of transactions with the help of various Unsupervised Machine Learning Algorithms. They use data mining techniques to spot patterns in a transaction that seem out of place. Credit Card Fraud Detection Frameworks are used so that the losses incurred by the merchants due to fraudulent transactions can be reduced and prevented.
* In my opinion, we put unlabeled features of different datapoints where we do not have access to the ground truth, such as time of the transaction, amount of money, location of the transaction, etc., and the anomaly is detected if any of these features show abnormal behaviour.



Based on the above diagram from a source (put all my references at the end of this document), when a transaction is made using a credit card, there is a Terminal Check that occurs - Is the pin correct? Does the person have sufficient balance? Is it a blocked card?

1. After the acceptance of all of these requirements, the transaction is fed as an observation into the predictive ML model.
2. If the Fraud Score is high or an anomaly is detected by the model, the transaction is then checked by investigators.
3. They confirm if it is actually a mistake, or the transaction is actually a fraud. The feedback is then fed into the model and the entire steps are repeated.

Machine Learning algorithms would need to analyze millions of data points that tend to go unnoticed by humans. Further, ML also reduces the number of false rejections and helps improve the precision of real-time approvals. These models are generally built on the client’s behaviour on the internet and transaction history.

Apart from spotting fraudulent behaviour with high accuracy, ML-powered technology is also equipped to identify suspicious account behaviour and prevent fraud in real-time instead of detecting them after the crime has already been committed, based on recent sources.

* *Data Requirements And Challenges:*
  + We need to gather data over a long period of time so that we can safely ensure the reliability of the machine learning model, so that it can understand patterns better. More data means more results and better patterns identified from unlabeled examples. We can source this out from bank transactions records data or published documents online such as Statista or company database for example.
  + A huge amount of data needs to be processed by the machine learning model in real-time and in very limited time frame, so we need to generate a robust model which can incorporate a healthy amount of training and test data and ensure that the model neither overfits nor underfits.
  + We need to train data from various historical transactions such as payments made by bank customers made over a yearly period overall, transaction parameters, and consumer behavior patterns in real-time to detect security threats and potential fraud cases.
  + The datasets used for training the models are usually imbalanced as most of the training data available consists of non-fraudulent transactions, which makes the detection of fraudulent ones extremely difficult.

These challenges can be overcome by building quick and straightforward machine learning models to detect anomalies and classify the transactions correctly. The highly imbalanced dataset can be sampled in a hybrid approach where the positive class is oversampled and the negative class under-sampled, achieving two sets of data distributions which can then be used as the training dataset. Machine learning algorithms like K nearest neighbours, Random Forest algorithm, Decision trees could be used to build the classification model, and to use skLearn metrics like accuracy score, precision, recall, and confusion matrix to validate and compare the performance of these classification models.

* *Security And Privacy Considerations:*
* **Biased data** — all supervised ML algorithms assume that the future would follow similar patterns as the past. The result is that models can make predictions based on underlying human characteristics unrelated to one’s ability to repay a loan, also called “bias against groups or classes of people”.
* **Regulations** — There is an act called “The Fair Credit Reporting Act (FCRA)” which aims to maintain fairness in credit reporting and ensure consumer’s privacy by protecting or using certain information and banks would face serious consequences if the data is leaked to the public. However, it is considered very unlikely that deep learning algorithms which require thousands of data points comply with FCRA. Creditors which rely on big data to make decisions often circumvent these regulations by denying revealing any model’s specifications, stating they are a “trade secret”. This is not a sustainable solution, and both businesses and researchers are exploring alternative options.

To achieve higher security levels and to prevent data leakage, dedicated algorithms analyze cardholders’ transaction data and other actions. Such algorithms can detect suspicious activity that potentially threatens account holders, increasing customers’ assets security. Payment systems like Payoneer and government agencies actively use these technologies to prevent money laundering. Machine learning in banking can evaluate enormous data sets of simultaneous transactions in real time.

* *Value Proposition:*

To ensure that financial institutions are able to manipulate and use machine learning algorithms to ensure that company’s transactions systems are kept robust, with key documentations in place, ensure a safe banking environment, eliminate suspicious activities and boost business profitability in the long run, as money laundering can cost businesses millions (such as the case for Deutsche Bank where they lost millions in a money laundering incident in the past).

1. Redo Problem 1 for the reinforcement learning scenario. Give one example of an application scenario that is different from the examples discussed in the lectures.

**ANSWER 3:**

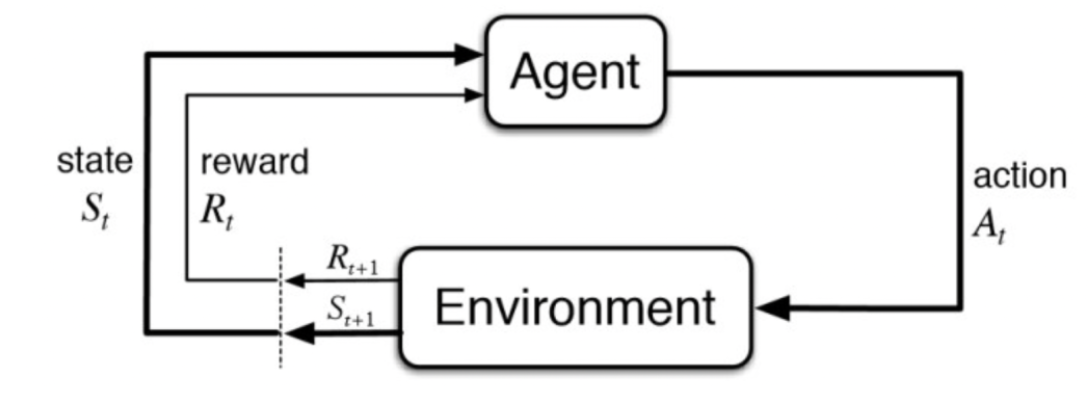
* *Industry Chosen:*

**Industry 4.0**

* *Technical Problem To Be Solved:*

**Applications Of Reinforcement Machine Learning On Self-Driving (Autonomous) Vehicles**

* *Working Methodology:* 
  + **Reinforcement learning (RL)** is a type of machine learning where an agent learns by exploring and interacting with the environment.
  + In the case of self-driving (autonomous) cars, the self-driving car is an agent.
  + In reinforcement learning, you can design an environment and use a reinforcement learning algorithm to optimize the driving policy. With a lot of learning, you can even work with multiple agents that explore multiple paths at the same time and returns you the optimal one.
  + In the case of autonomous vehicles, instead of learning from a labeled dataset (or unlabeled), you learn from mistakes generated by a reward system. We can say the agent i.e. the driving car learns from experience. In this technique, an agent (the car) interacts with an environment (the world) to reach an optimal state or objective. To interact, the agent performs an action.
  + Deep *Q*-Learning algorithm to control simulated cars, end-to-end, autonomously. The algorithm is based on reinforcement learning which teaches machines what to do through interactions with the environment.
  + If the action is positive, the reward will be big. If the action is negative such as going on the sidewalk and hurting people, the reward is negative. The goal of the agent is to choose the right action that gets the biggest reward.
  + The action taken is transitioned into some new state and the agent is given a reward. This process of evaluating a state, taking action, changing states, and rewarding is repeated. Throughout the process, it’s the agent’s goal to maximize the total amount of rewards.
* *Data Requirements And Challenges:*
* Autonomous vehicles make use of deep reinforcement learning (DRL) algorithms which have 3 important data requirements to take note of:
  1. **State** describes the current situation in a given time. In this case, it would be a position on the road.
  2. **Action** describes all the possible moves that the car can make.
  3. **Reward** is feedback that the car receives whenever it takes a certain action.



* We can get the data from the car position over time and all the actions that the car makes using simulation based software where we can make use of their databases to extract the essential data needed for training into the DRL.
* In DRL, the algorithm learns by exploring the environment and each interaction yields a certain reward. The reward can be both positive and negative. The goal of the DRL is to maximize the cumulative rewards.
* The data needed for reinforcement learning should be trained on perception data, where it learns what decision it should make which is a Convolutional Neural Network (CNN).
* During training, the agent (the car) learns by taking a certain action in a certain state. Based on this **state-action** pair, it receives a **reward.** This process happens over and over again. Each time the agent updates its memory of rewards. This is called the **policy. The policy is described as how the agent makes decisions.**It’s a decision-making rule. The policy defines the behaviour of the agent at a given time.
* In terms of challenges in obtaining the data, we would need to work with multiple agents i.e., cars which would augment the data required for training, meaning that we might need multiple models to analyze these large data gathered from the different positions, feedback and responses from cars motion which might be expensive, and would require extensive simulations.
* In self-driving cars, there are various aspects to consider, such as speed limits at various places, drivable zones, avoiding collisions for example, and these precautions that need to be taken might be a challenge in the data acquisition process.
* *Security And Privacy Considerations:*
* In my opinion, there are security concerns for other road users when sourcing out data from self-driving cars when data leakage occurs:
  1. Collision risk with other cars and passengers – passenger safety concern
  2. Data leakage from self-driving cars might lead to malfunction and pose a safety hazard to other vehicles in the vicinity
  3. Data leakage from user data i.e., the driver in the self-driving car would lead to a privacy breach as self-driving cars contain massive amounts of user data which could be a potential target for cybercriminals where they would have the power to control and deliberately alter the functionality of the autonomous vehicle such as car jamming, blinding, spoofing and abrupt braking which would pose a huge safety risk to the user involved.
  4. Self-driving cars could be hacked with ransomware, not allowing owners to enter, start or exit the vehicle until a ransom is paid.
  5. Terrorist hackers could disable networks, range sensors and cameras, resulting in multiple collisions.
  6. An autonomous vehicle’s operating system could be hacked, exposing personal information on other connected devices.
  7. There might even be situations where connected cars can control IoT devices at home, giving hackers access to people’s home computer networks.
* *Value Proposition:*
* In my opinion, the biggest value proposition that I can think of in this situation is that Machine Learning applications in the development of smart, connected and autonomous vehicles would be as follows:
  1. Accelerate Smart Nation development which is high on the government’s agenda.
  2. Boost digitization efforts which would lead to a more convenient 21st century for Gen Z individuals like myself where we do not need to take up the effort to drive.
  3. Job creation for millions of individuals, where this industry is a million dollar industry.

1. Suppose we want to remove vowels from a sentence. Write Python code to do this using iterators and list comprehension. The input sentence is: "The quick brown fox jumps over the lazy dog".
   1. Paste your Python code below, as well as the output of your program with the given input.

**ANSWER 4: a)**

**#variation 1 of vowel removal using list comprehension**

def new\_soln(sentence):

return "".join([char for char in sentence if char not in "aeiouAEIOU"])

sentence = "The quick brown fox jumps over the lazy dog"

print("Filtered result: " + new\_soln(sentence))

Output: Filtered result: Th qck brwn fx jmps vr th lzy dg

**#variation 2 of vowel removal using list comprehension**

def new\_soln (sentence):

vowels = 'aeiouAEIOU'

return "".join([ i for i in sentence if i not in vowels])

sentence = "The quick brown fox jumps over the lazy dog"

print("Filtered result: " + new\_soln(sentence))

Output: Filtered result: Th qck brwn fx jmps vr th lzy dg

**#variation 3 of vowel removal using iterators**

def iterator\_type(sentence):

vowels = 'aeiouAEIOU'

filtered\_list = []

for i in sentence:

if i not in vowels:

filtered\_list.append(i)

return ''.join(filtered\_list)

sentence = "The quick brown fox jumps over the lazy dog"

print("Filtered result: " + iterator\_type(sentence))

Output: Filtered result: Th qck brwn fx jmps vr th lzy dg

* 1. Submit your iPython notebook file (ipynb file) as well as a pdf print preview of the ipynb file as instructed above.

**ANSWER 4: b)**

Attached in the zip folder

**References**

1. <https://www.linkedin.com/pulse/machine-learning-use-cases-banking-finance-insurance-medium-gade/>
2. <https://www.itransition.com/machine-learning/banking>
3. <https://kindgeek.com/blog/post/5-top-machine-learning-use-cases-in-finance-and-banking-industry#text2>
4. <https://neptune.ai/blog/self-driving-cars-with-convolutional-neural-networks-cnn#:~:text=of%20processing%20steps.-,Reinforcement%20learning%20used%20for%20self%2Ddriving%20cars,driving%20car%20is%20an%20agent>.
5. <https://www.thinkautonomous.ai/blog/deep-reinforcement-learning-for-self-driving-cars-an-intro/>
6. <https://link.springer.com/chapter/10.1007/978-981-15-1081-6_5>
7. <https://eatron.com/reinforcement-learning-in-real-world-autonomous-driving-part-4/>
8. <https://neptune.ai/blog/reinforcement-learning-applications>
9. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8505783/>
10. <https://softengi.com/projects/healthcare-pharmaceutics/supervised-machine-learning-for-healthcare-industry-bacteria-identification-classification/>
11. <https://academic.oup.com/jacamr/article/3/1/dlab002/6127116>