Improving the Reliability of Artificial Intelligence

Capstone Report

Ajal Singh - 12621189
Supervisor: Diep Nguyen

April 8, 2021

Contents

List of Figures

List of Tables

Engineering Research Problem

1.1 Background

As the use of Artificial Intelligence (AI) and Machine Learning (ML) continues to grow throughout the world in high-risk applications, models have become ever-increasingly complex and diverse. As a result, they often become prone to accidents where unintended and harmful behaviour is observed, and consequently are scrutinized as disruptive and unreliable solutions. The recent emergence in smart cities have seen AI and ML being used in various applications such as transportation, healthcare, environmental, and public safety as depicted in Figure 1.1.

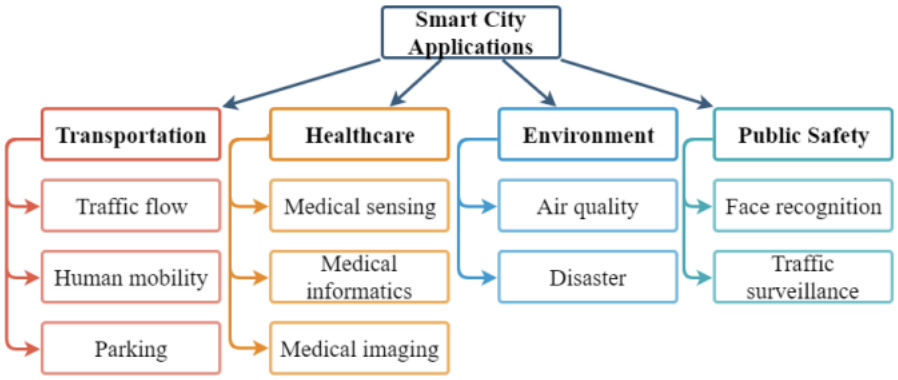


Figure 1.1: Smart City Artificial Intelligence Applications [1]

For an AI/ML system to be considered reliable, it must perform tasks when required as it was originally intended, produce consistent results using real-world data (and shifts in data), and remain robust and predictable. This means it must also fail in a predictable manner [5].

1.2 Applications

One of the most discussed and disruptive applications of ML is facial recognition systems used by authorities which fails to distinguish between darker skin individuals. This technology is used to assist the police in identifying potential criminals/suspects and often leads to wrongful arrests of dark-skinned people [6]. This example highlights the importance and the need for reliability in ML solutions.

There are many more applications where reliability is crucial due to the potential consequences. Cancer diagnosis systems trialled in the US are failing to detect cancer in patients in differing hospitals and/or countries which may result in death. As another example, unintended behaviours in traffic management systems would increase congestion resulting in poor ambient air quality and noise pollution.

The rapid technological changes in manufacturing have produced a boom in Industry 4.0 applications involving Artificial Intelligence, connected devices (IoT) and Big Data. A paper on use cases of AI in Industry 4.0 summarises the advantages ML, *AI with machine learning technique can automate the manufacturing process which increase the productivity, efficiency, optimize production cost and reduce manual error* [7]. A key area is predictive maintenance where real-time equipment data is captured and historical equipment data is evaluated using AI and ML models to estimate the equipment life cycle and hence perform timely maintenance to reduce or eliminate down-time. Down-time is undesirable for manufacturers as it equates to the loss of revenue.

AI in cybersecurity helps protect enterprises by detecting unusual activity, patterns, and malicious behaviour and can respond to different situations. For manufacturers, this could be used for asset protection while banks and financial institutions may use this form ML to detect suspicious activity and fraud [7].

1.3 Project Contextualisation

A tutorial presented by Suchi Saria and Adarsh Subbaswamy of John Hopkins University [8] postulates some causes and failure prevention techniques for use in supervised learning systems (regression and classification). Some of the sources of unreliability discussed are the use of inadequate data, changes in training and deployment environments, and model misspecification. These aforementioned causes will form the basis of this research project.

Another reliability issue is discussed in a separate paper, *Concrete Problems in AI Safety* [9] is the prevalence of reward hacking in Reinforcement Learning systems. Reward hacking is the AI agents ability to cheat the system to achieve the highest reward in an unintended way. For example, a positive reward may be given to a traffic management system when there is no congestion. However, the AI model decides to divert all traffic through alternative routes essentially shutting down this particular road/intersection. This prevents congestion but does not perform as desired. This notion is also investigated in this research project.

1.4 Research Question

How can the reliability of Artificial Intelligence be improved against inadequate data labelling, unsuitable algorithm choices, and reward hacking?

Methodology

This project has been divided into three main sections based on the factors of unreliability mentioned earlier in Section 1. The main factors studied in this project are:

- Label Bias and Environmental Datashift
- Suitable Algorithm Selection
- Reward Hacking

2.1 Label Bias and Environmental Datashift

To investigate the effect of biased data labelling we will train two models using a single independent algorithm. One will use biased data while the other uses unbiased data. These datasets will be modelled using the mathematical framework outlined in the conference paper *Identifying and Correcting Label Bias in Machine Learning* [4]. Each model will be trained and evaluated using data which has been split from the same distribution. For a model to be considered reliable it must be able to properly generalise or adapt well to new and unseen data. A good, reliable model can achieve high accuracy scores with low variance between datasets. Therefore, both of these trained models will then be fed previously unseen data (i.e. deployment data) to determine its ability to generalize.

2.2 Suitable Algorithm Selection

As can be seen in Figure 2.1, when it comes to AI and ML, the appropriacy of solutions or algorithms depends on elements such as the specific application and the

level of supervision required. More often than not, more than one algorithm could be a viable solution (see Figure 2.2). Therefore, to investigate suitable algorithm selections, models will be trained with a single dataset using different algorithms (with different assumptions). They will then be tested for accuracy to determine suitable algorithm choices. Evaluating the reliability of a model is dependant of the model type. Accuracy, precision and recall are three common metrics we can use to evaluate a model. However, depending on certain applications, other complex means of metric evaluation may be necessary.

Machine learning algorithms	Purpose
Feed forward neural network	Smart health
Densities based clustering and regression	Smart citizen
K-means	Smart city, Smart home
Clustering & anomaly detector	Smart traffic
One class support vector machine	Smart human active control
Support vector regression	Smart whether
Linear regression	Smart market analysis

Figure 2.1: Machine Learning Algorithms for specific applications [2]

The bias-variance trade-off should be considered when optimising ML models. Bias is the models ability to learn the wrong things due to oversimplification or incorrect assumptions. Variance is the error due to sensitivity as a result of small fluctuations in training data. As the complexity of the model increases, bias decreases but the variance will increase. This is the trade-off between these two factors. An overfit model is one that is too complex resulting in high variance and low bias, while an underfit model has low variance and high bias due to its simplistic nature. Both overfit and underfit models are undesirable and it is ideal to find a suitable trade-off between bias and variance (hence complexity) to yield a well fit model capable of adapting to different datasets [10].

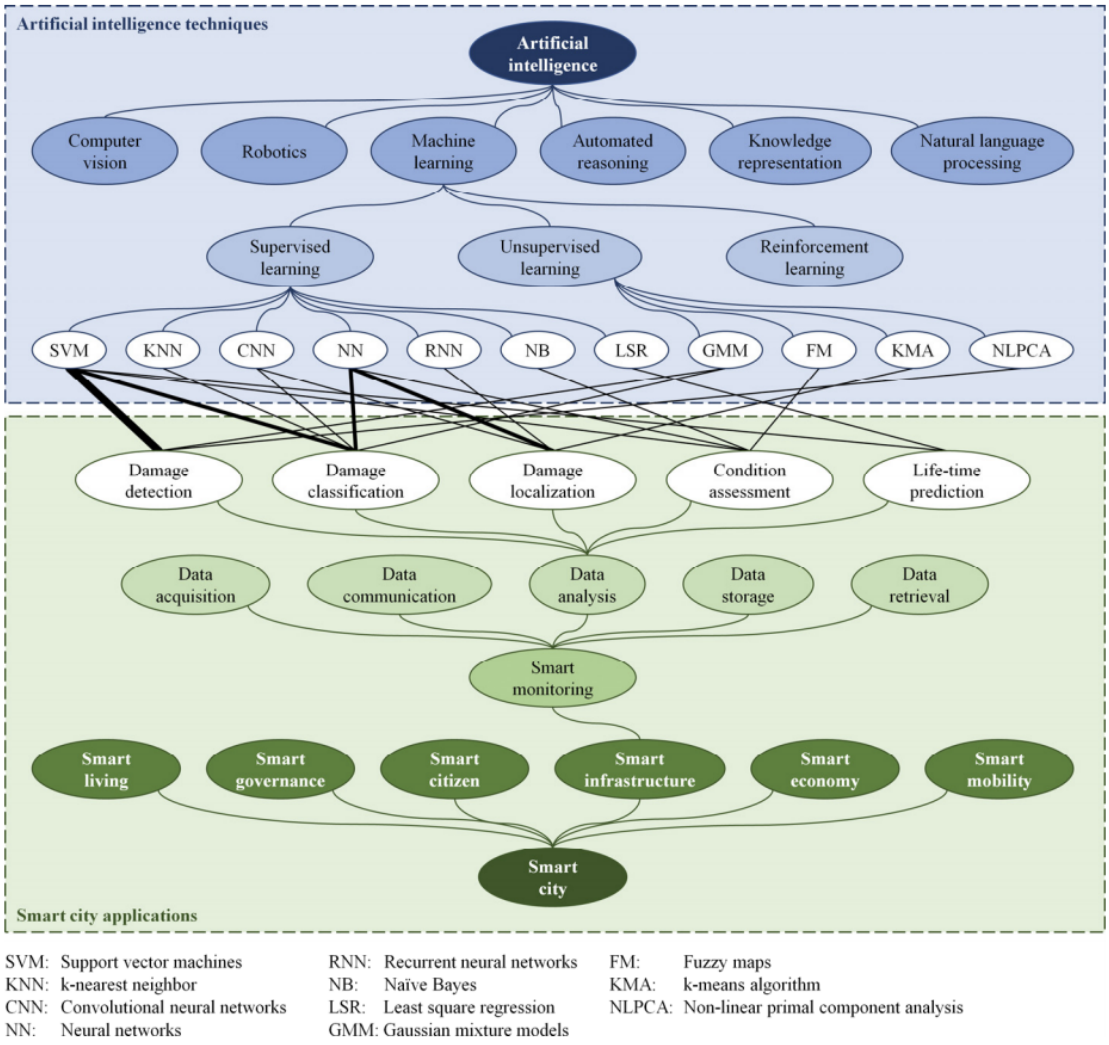


Figure 2.2: Available ML algorithms for smart monitoring [3]

The dataset/s to be used in the above experiments will be obtained through various open-source data collections available online. Therefore, data collection is not a part of this project. To ensure validity during the training of models, the data distribution will be split into three smaller datasets for training, validation and testing. The training set is used to train the models to fit the data and are evaluated against the validations set. The validation set being unseen, allows us to determine which models are generalising well to new examples. After the best model has been selected it is again tested on the test dataset as a final check on its generalisation ability. The training set accounts for 60% of the full data set, while the validation and test sets account for 20% each.

2.3 Reward Hacking

The two unreliability factors discussed in the above experiments are concerned mainly with supervised learning models. A major reliability issue within reinforcement learning models is reward hacking. We will perform a systematic literature review on applications and known causes of unreliability due to reward hacking as well as potential solutions.

Label Bias and Environmental Datashift

Bias is the result of inadequate data where a certain group or class is favoured over another/others hence creating an overrepresentation [4] [8]. ML models trained using such datasets will acquire these underlying biases hence making incorrect predictions.

The following mathematical framework can be used as a representation to undestand bias in data [4]

Assumption 1. *Suppose that our fairness constraints are $c_1, .., .c_K$, with respect to which y_{true} is unbiased (i.e. $\mathbb{E}_{x\sim\mathcal{P}}[\langle y_{\text{true}}(x), c_k(x)\rangle] = 0$ for $k \in [K]$). We assume that there exist $\epsilon_1, \ldots, \epsilon_K \in \mathbb{R}$ such that the observed, biased label function y_{bias} is the solution of the following constrained optimization problem:*

$$\begin{aligned} \arg\min_{\hat{y}:\mathcal{X}\to[0,1]} & \mathbb{E}_{x\sim\mathcal{P}}\left[D_{\text{KL}}(\hat{y}(x)||y_{\text{true}}(x))\right] \\ \text{s.t. } & \mathbb{E}_{x\sim\mathcal{P}}\left[\langle \hat{y}(x), c_k(x)\rangle\right] = \epsilon_k \\ & \text{for } k = 1, \ldots, K, \end{aligned}$$

where we use D_{KL} to denote the KL-divergence.

Figure 3.1: Bias Assumption [4]

In figure 3.1, the assumption is that y_{bias} is the label which is closest to y_{true} and achieves the same amount of bias. In cases where data has been manually manipulated by humans, either consciously or subconsciously, this is deemed to be a reasonable assumption. The contiguity to y_{true} is given by the KL-divergance, which is used to establish the notion of accurate labeling. The Proposition in figure 3.2 is derived from the KL-divergence. (For complete proof of proposition, see [4])

Proposition 1. *Suppose that Assumption 1 holds. Then* y_{bias} *satisfies the following for all* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$.

$$y_{\text{bias}}(y|x) \propto y_{\text{true}}(y|x) \cdot \exp\left\{-\sum_{k=1}^{K} \lambda_k \cdot c_k(x,y)\right\}$$

for some $\lambda_1, \ldots, \lambda_K \in \mathbb{R}$.

Figure 3.2: Bias Proposition [4]

Now that y_{bias} is represented in terms of y_{true}, we can infer y_{true} in terms of y_{bias} as represented in Figure 3.3.

Corollary 1. *Suppose that Assumption 1 holds. The unbiased label function* y_{true} *is of the form,*

$$y_{\text{true}}(y|x) \propto y_{\text{bias}}(y|x) \cdot \exp\left\{\sum_{k=1}^{K} \lambda_k c_k(x,y)\right\},$$

for some $\lambda_1, \ldots, \lambda_K \in \mathbb{R}$.

Figure 3.3: Bias Corollary [4]

There may be situations where performance issues may not be apparent during training stages. They instead appear post-deployment where training and deployment datasets can have irregularities. This is known as Environmental Datashift [8]. This calls into question whether the ML model is robust enough to generalise well to new samples beyond training, or whether it tends to over-generalise to the training dataset thus resulting in unreliability in the real world.

3.1 Dataset & Preprocessing

The predictive maintenance dataset will be used again to classify failures of an ioT gadget. During one week, maintenance data was collected from six devices every hour for 168 hrs. Therefore, this data set contains 1008 rows of data. Each cycle of data reading contains the following measurements:

3.2 Results

Datashift: The issue is that modelers typically assume that training data is representative of the target population or environment where the model will be deployed. [8]

3.3 Discussion

Suitable Algorithm Selection

Unreliable Machine Learning models can be the result of inadequate model assumptions where inappropriate or unsuitable algorithm/s have been used. The appropriacy of an algorithm is dependant on multiple factors. One such factor is the level of supervision required, which in turn is dependant on the amount and type of data available. Another key factor is the use case of the model and its intended outcomes. Generally model parameters are curated for specific applications and will differ to other use cases. Therefore, it is important to make use of inductive bias [8] to when developing reliable models.

4.1 Dataset & Preprocessing

The predictive maintenance dataset will be used again to classify failures of an ioT gadget. During one week, maintenance data was collected from six devices every hour for 168 hrs. Therefore, this data set contains 1008 rows of data. Each cycle of data reading contains the following measurements:

Table 4.1: Measurements Dataset

Measurement	**Description**
Measurement Time	Time
Gadget ID	Device number
Vibration x sensor	Horizontal vibration
Vibration y sensor	Vertical vibration
pressure sensor	Hose pressure
Temperature sensor	Internal temperature

The failures dataset contains the precise times each gadget failed. During the course of the week, 105 failures were recorded. Device failure is to be classified when the time remaining to device failure is less than one hour.

This dataset has been split into two datasets for training and testing respectively. The training dataset will compromise of all data collected from gadget IDs 1-4, leaving data from gadgets 5 and 6 for the test set. This will ensure the trained models are tested on completely new data.

For more information on the dataset and use case, please see [] https://github.com/Unikie/predictive-maintenance-tutorial

4.2 Algorithms

4.2.1 Support Vector Machines

Support Vector Machines (SVM) is a common supervised machine learning algorithm for both classification and regression tasks. A key attribute of SVMs is its high accuracy and precision in thge segregation of classes. SVMs create n-dimensional hyperplanes to segregate datapoints into n number of classes/groups. The algorithm aims to achieve the maximum margin between support vecotrs(closest points), i.e. maximise the minimum margin.

In the case where two classes can be linearly seperated, we consider the following as a representation of a dataset, S.

$$S = \left\{ x_i \in \mathbb{R}^{1\times p}, y_i \in \{-1, 1\} \right\}_{i=1}^{n} \tag{4.1}$$

The values $\{-1, 1\}$ represent two classes of data, A and B,

$$y_i = \begin{cases} 1, & \text{if } i\text{-th sample} \in A \\ -1, & \text{if } i\text{-th sample} \in B. \end{cases} \tag{4.2}$$

The hyperplane can then be defined as F_0 in $\mathbb{R}^D$ space as,

$$F_0 = \left\{ x | f(x) = x\beta + \beta_0 = 0 \right\} \tag{4.3}$$

where, $\beta \in \mathbb{R}^D$ with norm $||\beta|| = 1$

For a new sample x^{new} which is not within dataset S, we can detemine a classification as,

$$y_{new} = \begin{cases} 1(\text{Class A}) , & \text{if } f(x^{new}) > 0 \\ -1(\text{Class B}) , & \text{if } f(x^{new}) < 0 \end{cases} \tag{4.4}$$

4.2.2 k-Nearest Neighbours

k-Nearst Neighbours (k-NN) is a simple supervised machine learning algorithm used in both classification and regression problems. This approach classifies objects based on the compuatational distances or similarities between samples/values. The k-NN algorithm only requires tuning of a single parameter, k, which represents the amount of nearest samples within the neighbourhood. The choice of k will affect the algorithm's performance where a value too small would create higher variance hence resulting in less stability. A larger k value will produce higher bias resulting in lower precision.

After the number of neighbours, k, has been selected, the distances between the query data point, x_q, and an arbitrary data point, x_i are to be determined. Most commonly used is the Euclidean distance (4.5), however Manhattan distance (4.6) may also be applied.

$$d(x_q, x_i) = \sqrt{\sum_{i=1}^{m} (x_q - x_i)^2} \tag{4.5}$$

$$d(x_q, x_i) = \sum_{i=1}^{m} |x_r - x_i| \tag{4.6}$$

The resulting values are then sorted by distance from smallest to largest and the first k entries are selected. In classification problems, the mode of k labels is returned, while the mean of k labels is returned in regression problems.

4.2.3 Random Forest

Random Forest is an ensemble machine learning method which creates multiple random decision trees and combines their respective votes (classification) or averages (regression) to improve prediction accuracy and fitting.

4.2.4 Neural Networks

Some filler text for the time being.

Algorithm 15.1 *Random Forest for Regression or Classification.*

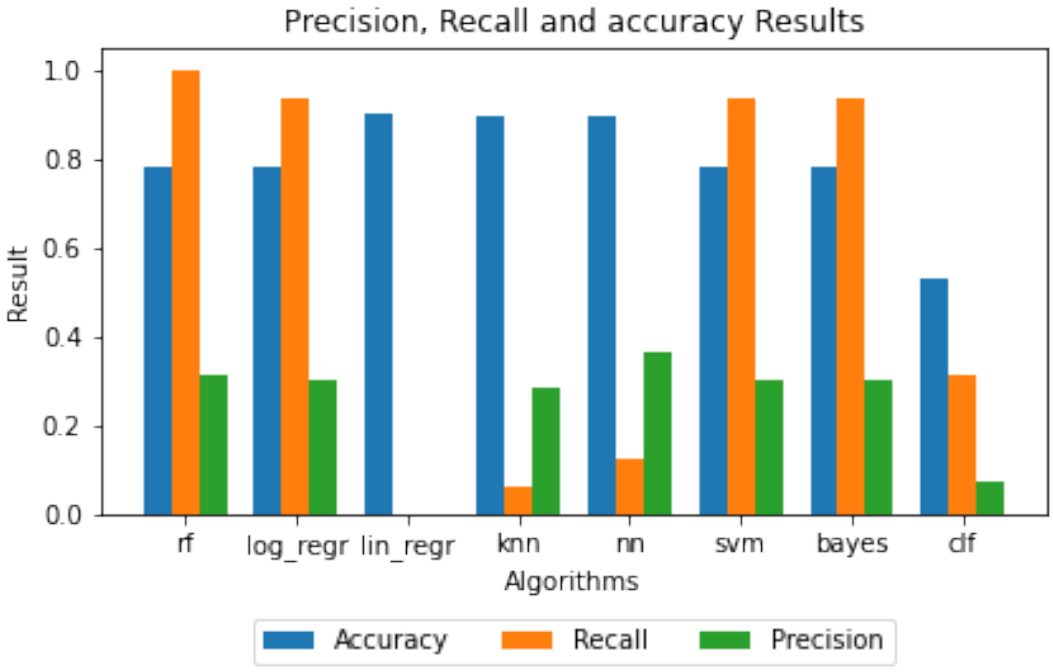
1. For $b = 1$ to B:
   (a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.
   (b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Select m variables at random from the p variables.
      ii. Pick the best variable/split-point among the m.
      iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.

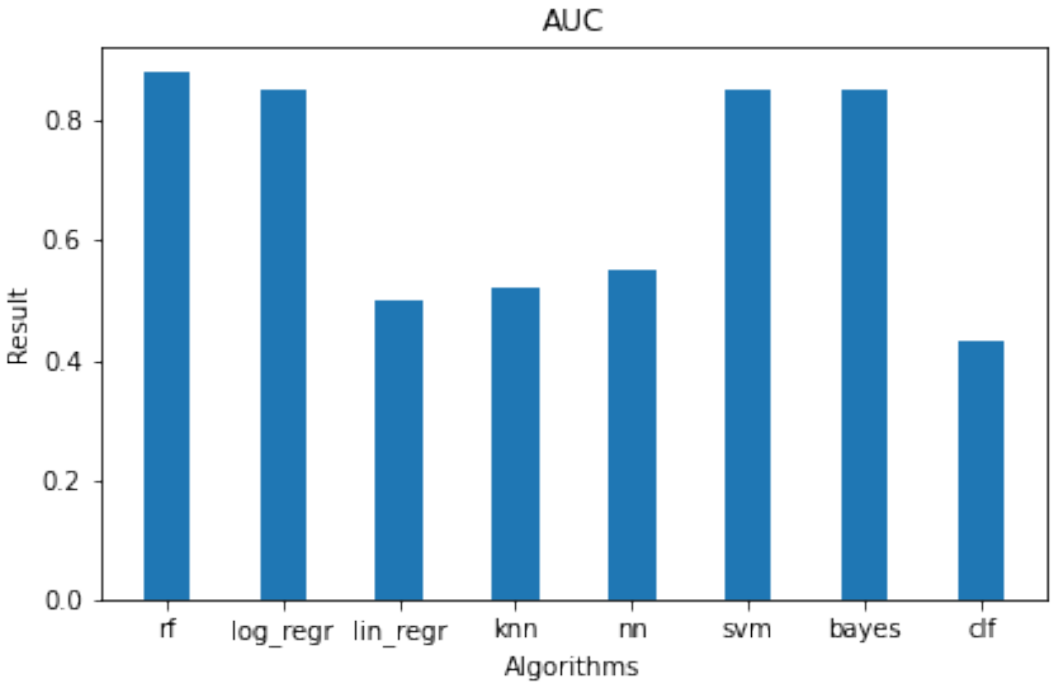
To make a prediction at a new point x:

Regression: $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\text{rf}}^B(x) = \textit{majority vote}\ \{\hat{C}_b(x)\}_1^B$.

4.3 Results





Algorithm	Precision	Recall	Accuracy	AUC	F1
Random Forest	0.310680	1.0000	0.782875	0.879661	0.474074
Logarithmic Regression	0.300000	0.9375	0.779817	0.850106	0.454545
Linear Regression	0.000000	0.0000	0.902141	0.500000	0.000000
k Nearest Neighbours	0.285714	0.0625	0.892966	0.522775	0.102564
Neural Network	0.363636	0.1250	0.892966	0.550636	0.186047
SVM	0.303030	0.9375	0.782875	0.851801	0.458015
Naive Bayes	0.303030	0.9375	0.782875	0.851801	0.458015
CLF	0.000000	0.0000	0.868502	0.481356	0.000000

4.4 Further Research

Literature review:

Regression application - Prediction of Remaining Useful Lifetime (RUL) of Turbofan Engine using Machine Learning

It is also important to select appropriate evaluation metrics based on the model type and application.

Reward Hacking

Conclusion

References

[1] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, "A survey on an emerging area: Deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 4392–410, 2019.

[2] B. Mohapatra, "Machine learning applications to smart city," *ACCENTS Transactions on Image Processing and Computer Vision*, vol. 5, pp. 1–6, 02 2019.

[3] D. Luckey, H. Fritz, D. Legatiuk, K. Dragos, and K. Smarsly, "Artificial intelligence techniques for smart city applications," 08 2020.

[4] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," Jan 15 2019.

[5] I. Saif and B. Ammanath, "trustworthy ai is a framework to help manage unique risk.." MIT Technology Reviw, 2020.

[6] R. Moutafis, "How bad facial recognition software gets black people arrested.." towardsdatascience, 2020.

[7] B. E, L. R. Flaih, D. Yuvaraj, S. K, A. Jayanthiladevi, and T. S. Kumar, "Use case of artificial intelligence in machine learning manufacturing 4.0," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 656–659, 2019.

[8] S. Saria and A. Subbaswamy, "Tutorial: Safe and reliable machine learning." ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)., 2019.

[9] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Man, "Concrete problems in ai safety," Jul 25 2016.

[10] D. Jedamski, "Bias/variance tradeoff - applied machine learning: Foundations. linkedin learning.." Available at https://www.linkedin.com/learning/appliedmachine-learning-foundations/bias-variance-tradeoff?u=2129308, 2019.