

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KỸ THUẬT ĐỊA CHẤT VÀ DẦU KHÍ**



LUẬN VĂN TỐT NGHIỆP

**ỨNG DỤNG CÁC PHƯƠNG PHÁP HỌC MÁY
TRONG DỰ BÁO KHAI THÁC DẦU KHÍ**

**APPLYING MACHINE LEARNING METHODS
FOR PRODUCTION RATE PREDICTION**

Sinh viên : Huỳnh Nguyễn Hiếu Nghĩa
MSSV : 1712314
GVHD : ThS. Trần Nguyễn Thịện Tâm

TP. HCM, 12/2021

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
Số: /ĐHBK-ĐT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

KHOA: KỸ THUẬT ĐỊA CHẤT VÀ DẦU KHÍ
HỌ VÀ TÊN: HUỲNH NGUYỄN HIẾU NGHĨA
NGÀNH: KHOAN – KHAI THÁC DẦU KHÍ

MSSV: 1713970
LỚP: DC17K

1. Đề tài luận văn:

NHIỆM VỤ CỦA LUẬN VĂN TỐT NGHIỆP

ỨNG DỤNG CÁC PHƯƠNG PHÁP HỌC MÁY TRONG DỰ BÁO KHAI THÁC DẦU KHÍ

APPLYING MACHINE LEARNING METHODS FOR PRODUCTION RATE PREDICTION

2. Nhiệm vụ của luận văn tốt nghiệp:

- Nghiên cứu tổng quan về các mô hình dự báo khai thác dầu khí thông thường.
- Nghiên cứu tổng quan về các thuật toán học máy và học sâu.
- Dự báo lưu lượng dầu khí khai thác giếng 15/9-F-14, mỏ Volve với các thuật toán học máy và học sâu và ngôn ngữ lập trình python.
- Đánh giá các kết quả đạt được.

3. Ngày giao nhiệm vụ luận văn: / /2021

4. Ngày hoàn thành nhiệm vụ luận văn: / /2021

5. Họ và tên người hướng dẫn: Phàn hướng dẫn:
ThS. Trần Nguyễn Thiện Tâm Toàn bộ luận văn

Nội dung và yêu cầu LVTN đã thông qua Bộ môn Khoan – Khai thác Dầu khí thuộc Khoa Kỹ thuật Địa chất và Dầu Khí

Ngày.....tháng.....năm 2021

CHỦ NHIỆM BỘ MÔN
(Ký và ghi rõ họ tên)

CÁN BỘ HƯỚNG DẪN CHÍNH
(Ký và ghi rõ họ tên)

TS. Mai Cao Lân

ThS. Trần Nguyễn Thiện Tâm

PHẦN DÀNH CHO KHOA, BỘ MÔN:

Người duyệt (chấm sơ bộ):

Đơn vị:

Ngày bảo vệ:

Điểm tổng kết:

Nơi lưu trữ luận văn:

LỜI CẢM ƠN

Từ cổ chí kim, từ Đông sang Tây, dù là trong bất kỳ nền văn minh nào, việc học vẫn luôn được xem trọng. Không Tử đã từng nói rằng: “Biết thì nói là biết. Không biết thì nói là không biết. Thέ mới gọi là biết”. Hay như nhà hoạt động chống chủ nghĩa phân biệt chủng tộc Apartheid – Nelson Mandela từng nhận định: “Giáo dục là vũ khí mạnh nhất mà người ta có thể sử dụng để thay đổi cả thế giới”. Có thể nói việc học không chỉ là con đường khai mở tâm trí, mà còn sợi chỉ xâu chuỗi mọi tri thức của nhân loại và là chìa khóa mở ra cánh cửa dẫn đến tương lai.

Là một người học, em ý thức được tầm quan trọng của việc học và học không ngừng. Chặng đường đồng hành với trường đại học Bách Khoa và khoa Kỹ thuật Địa chất & Dầu khí không chỉ mang lại cho em những tri thức cần thiết đối với ngành nghề, mà còn mở ra cho em nhiều góc nhìn khác nhau trong cách tiếp cận mọi vấn đề. Qua đó, nó khuyến khích em phải học nhiều hơn nữa, tiếp tục khám phá những tri thức mới.

Kết thúc một chặng đường học tập, em muốn gửi lời cảm ơn chân thành đến quý thầy cô khoa Kỹ thuật Địa chất & Dầu khí, cũng như quý thầy cô trường đại học Bách Khoa nói chung. Đặc biệt, em xin chân thành cảm ơn thầy **Trần Nguyễn Thiện Tâm**, giảng viên bộ môn Khoan và Khai thác Dầu khí, chủ nhiệm lớp DC17KK. Cảm ơn thầy đã luôn sát cánh cùng lớp DC17KK và tận tâm hướng dẫn em trong việc hoàn thành luận văn tốt nghiệp. Bên cạnh đó, con xin gửi lời cảm ơn đặc biệt đến ba mẹ. Cảm ơn ba mẹ đã chịu thương chịu khó trong suốt quãng đường học tập của con, qua đó con muốn gửi món quà quý giá nhất đến cho ba mẹ chính là những thành quả mà con đạt được trong suốt quãng thời gian ấy. Cảm ơn những người bạn đã đồng hành cùng mình, cùng nhau vượt qua những năm tháng đại học nhiều thử thách, mà cũng lắm niềm vui.

Chặng đường học tập phía trước không còn dễ dàng như chặng đường học tập vừa qua. Tuy nhiên, những tri thức mà em góp nhặt được trong suốt những năm tháng đại học sẽ là những hành trang quý báu giúp em vượt qua mọi thử thách.

TÓM TẮT LUẬN VĂN

Việc phân tích khai thác từ lâu đã trở thành một nhiệm vụ quan trọng trong ngành dầu khí. Từ nhu cầu ước tính trữ lượng và đánh giá tiềm năng via đã đặt ra tiền đề cho việc nghiên cứu về các phương pháp phân tích và dự báo khai thác. Qua đó, nhiều cách tiếp cận vấn đề đã được đề xuất hàng thập kỷ qua, chẳng hạn như phương pháp phân tích đường cong suy giảm (DCA), type – curves, mạng nơ – ron nhân tạo (ANN) và gần đây là các phương pháp học máy và học sâu.

Luận văn nghiên cứu về chủ đề ứng dụng trí tuệ nhân tạo trong lĩnh vực dầu khí, mà cụ thể là ứng dụng các phương pháp học máy và học sâu dự báo sản lượng dầu khí khai thác trong tương lai. Bài toán cần giải quyết là một bài toán ngoại suy, do đó các phương pháp hồi quy được quan tâm xem xét.

Từ một bộ dữ liệu cho trước, nhiệm vụ chính là từ một bộ số liệu cho trước phải đưa ra một mô hình dự báo hiệu quả phục vụ cho quá trình khai thác. Về dữ liệu, với sự phát triển của dữ liệu lớn, ngày nay ta hoàn toàn có thể xử lý một tập dữ liệu cực lớn, có đầy đủ các thông số cần thiết cho quá trình nghiên cứu một cách dễ dàng và không mất thời gian. Bên cạnh đó, ứng dụng một số kỹ thuật trong lĩnh vực khoa học dữ liệu, luận văn đã tiến hành khảo sát sơ bộ, đánh giá và tinh chỉnh bộ dữ liệu gốc, làm nó phù hợp hơn với các phương pháp học máy và học sâu mà không làm mất đi những đặc điểm toán học vốn có của nó.

Có 3 phương pháp học máy hồi quy được khảo sát, gồm: hồi quy tuyến tính, máy véc – to hỗ trợ. Với học sâu, họ mạng nơ – ron hồi quy sẽ được khảo sát, gồm: mạng nơ – ron hồi quy đơn giản (Simple RNN), bộ nhớ ngắn hạn dài (LSTM) và nút hồi quy có công (GRU).

Các phương pháp học máy và học sâu được xây dựng bằng ngôn ngữ lập trình Python trên nền tảng Pycharm. Các thư viện hỗ trợ cho việc xây dựng mô hình có thể kể đến như sklearn, keras, matplotlib, pandas, numpy, ...

MỤC LỤC

NHIỆM VỤ CỦA LUẬN VĂN TỐT NGHIỆP.....	ii
LỜI CẢM ƠN.....	i
TÓM TẮT LUẬN VĂN.....	ii
MỤC LỤC	iii
DANH MỤC HÌNH ẢNH.....	vi
DANH MỤC BẢNG BIÊU	x
DANH MỤC BIỂU ĐỒ	xii
DANH MỤC KÝ HIỆU VIẾT TẮT VÀ THUẬT NGỮ	xiii
MỞ ĐẦU	xiv
1. Tính cấp thiết của đề tài	xiv
2. Mục tiêu nghiên cứu	xiv
3. Nhiệm vụ của luận văn	xv
4. Phương pháp nghiên cứu	xv
5. Ý nghĩa khoa học và thực tiễn	xv
6. Tổng quan tình hình nghiên cứu	xvi
7. Cấu trúc luận văn	xx
CHƯƠNG 1: TỔNG QUAN VỀ DỰ BÁO KHAI THÁC DẦU KHÍ	1
1.1 Lịch sử phát triển các phương pháp phân tích quá trình khai thác	1
1.2 Phân tích đường cong suy giảm Arps [6].....	4
CHƯƠNG 2: TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO	10
2.1 Giới thiệu chung	10

2.2 Kỹ thuật xây dựng đặc trưng [7]	12
2.2.1 Chuyển khoảng giá trị	13
2.2.2 Chuẩn hóa theo phân phối chuẩn	13
2.2.3 Chuẩn hóa về cùng norm.....	13
2.2.4 Xử lý dữ liệu rỗng	13
2.3 Các mô hình học máy hồi quy.....	14
2.3.1 Hồi quy tuyến tính (Linear Regression) [7]	14
2.3.2 Máy véc – tơ hỗ trợ (Support Vector Machine).....	17
2.4 Mạng nơ – ron hồi quy	24
2.4.1 Mạng nơ – ron hồi quy (Recurrent Neural Network) [9].....	24
2.4.2 Bộ nhớ ngắn hạn dài (Long Short – Term Memory) [9].....	27
2.4.3 Nút hồi quy có cổng (Gated Recurrent Unit) [10]	30
2.4.4 Các hàm kích hoạt thông dụng [11]	33
2.4.5 Các hàm tối ưu hóa thông dụng [12].....	34
2.5 Quá khớp	37
CHƯƠNG 3: DỰ BÁO KHAI THÁC GIÉNG F14 – MỎ VOLVE	39
3.1 Tổng quan về mỏ Vole	39
3.1.1 Mô tả mỏ	39
3.1.2 Lịch sử khai thác	41
3.2 Quy trình thực hiện.....	42
3.3 Dữ liệu đầu vào	43
3.4 Tiền xử lý dữ liệu	44
3.5 Xây dựng mô hình học máy hồi quy	51

3.6 Xây dựng mô hình mạng nơ – ron hồi quy	53
3.7 Đánh giá kết quả thực hiện.....	59
3.7.1 Đánh giá các mô hình học máy	59
3.7.2 Đánh giá các mô hình mạng nơ – ron hồi quy	67
3.7.3 So sánh các mô hình học máy với học sâu, học máy và học sâu với đường cong suy giảm DCA và mạng nơ – ron nhân tạo ANN	78
KẾT LUẬN VÀ KIẾN NGHỊ.....	81
1. Kết luận.....	81
2. Kiến nghị.....	82
TÀI LIỆU THAM KHẢO	83
PHỤ LỤC A	85
1. Quy trình phân tích đường cong suy giảm với mô hình Exponential.....	85
2. Quy trình phân tích đường cong suy giảm với mô hình Harmonic	86
3. Quy trình phân tích đường cong suy giảm với mô hình Hyberbolic	86
4. Dự báo lưu lượng dầu khí khai thác bằng phương pháp DCA	88
PHỤ LỤC B.....	100
1. Xây dựng mô hình mạng nơ – ron nhân tạo	100
2. Đánh giá kết quả dự báo khai thác dầu khí bằng mạng nơ – ron nhân tạo.....	104

DANH MỤC HÌNH ẢNH

Hình 1-1: Các loại đường cong suy giảm.....	6
Hình 2-1: Mối quan hệ giữa trí tuệ nhân tạo, học máy và học sâu.....	11
Hình 2-2: Mô hình học máy nói chung	12
Hình 2-3: Mô hình máy véc – tơ hỗ trợ (lè mềm)	18
Hình 2-4: Mô hình hồi quy vec – tơ hỗ trợ.....	22
Hình 2-5: Kiến trúc mạng nơ – ron hồi quy	25
Hình 2-6: Một ô nhớ LSTM điển hình	28
Hình 2-7: Nút hồi quy có cổng đàm đủ điển hình	31
Hình 2-8: Các hàm kích hoạt thường dùng	33
Hình 2-9: Các hiện tượng underfitting, fitting và overfitting.....	37
Hình 3-1: Vị trí của mỏ Volve.....	39
Hình 3-2: Vị trí mỏ Volve trong block 15/9 và các mỏ lân cận	40
Hình 3-3: Mô hình via của mỏ Volve.....	40
Hình 3-4: Dữ liệu khai thác trong giai đoạn 2008 – 2017	41
Hình 3-5: Quy trình chi tiết cho các thuật toán học máy và học sâu.....	43
Hình 3-6: Biểu đồ lưu lượng dầu khai thác trước khi dữ liệu được xử lý	47
Hình 3-7: Biểu đồ lưu lượng khí khai thác trước khi dữ liệu được xử lý	48
Hình 3-8: Biểu đồ lưu lượng khí khai thác trước khi dữ liệu được xử lý	48
Hình 3-9: Biểu đồ lưu lượng khí khai thác sau khi dữ liệu được xử lý	49
Hình 3-10: Ma trận tương quan giữa các đặc trưng	49
Hình 3-11: Lưu lượng dầu khai thác sau khi được tuyến tính hóa.....	51
Hình 3-12: Lưu lượng khí khai thác sau khi được tuyến tính hóa	52
Hình 3-13: Lưu lượng dầu sau khi gọi hàm	53
Hình 3-14: Lưu lượng khí sau khi gọi hàm	54
Hình 3-15: Cấu trúc họ mạng nơ – ron hồi quy.....	57
Hình 3-16: Dự báo sản lượng dầu khai thác theo mô hình hồi quy tuyến tính	60
Hình 3-17: Sản lượng dầu khai thác dự báo so với giá trị thực tế - Hồi quy tuyến tính ...	61
Hình 3-18: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy tuyến tính	61

Hình 3-19: Dự báo sản lượng khí khai thác theo mô hình hồi quy tuyến tính.....	62
Hình 3-20: Sản lượng khí khai thác dự báo so với giá trị thực tế – Hồi quy tuyến tính ...	62
Hình 3-21: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy tuyến tính	63
Hình 3-22: Dự báo sản lượng dầu khai thác theo mô hình hồi quy Ridge.....	64
Hình 3-23: Sản lượng dầu khai thác dự báo so với giá trị thực tế – Hồi quy Ridge	64
Hình 3-24: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy Ridge	65
Hình 3-25: Dự báo sản lượng khí khai thác theo mô hình hồi quy Ridge	65
Hình 3-26: Sản lượng khí khai thác dự báo so với giá trị thực tế – Hồi quy Ridge.....	66
Hình 3-27: Sản lượng khí khai thác tích lũy theo thời gian – Hồi quy Ridge	66
Hình 3-28: Dự báo sản lượng dầu khai thác theo mô hình Simple RNN.....	68
Hình 3-29: Sản lượng dầu khai thác dự báo so với giá trị thực tế – Simple RNN	68
Hình 3-30: Sản lượng dầu khai thác tích lũy theo thời gian – Simple RNN.....	69
Hình 3-31: Dự báo sản lượng khí khai thác theo mô hình Simple RNN	69
Hình 3-32: Sản lượng khí khai thác dự báo so với giá trị thực tế – Simple RNN.....	70
Hình 3-33: Sản lượng khí khai thác tích lũy theo thời gian – Simple RNN	70
Hình 3-34: Dự báo sản lượng dầu khai thác theo mô hình LSTM.....	71
Hình 3-35: Sản lượng dầu khai thác dự báo so với giá trị thực tế – LSTM	72
Hình 3-36: Sản lượng dầu khai thác tích lũy theo thời gian – LSTM	72
Hình 3-37: Dự báo sản lượng khí khai thác theo mô hình LSTM.....	73
Hình 3-38: Sản lượng khí khai thác dự báo so với giá trị thực tế – LSTM.....	73
Hình 3-39: Sản lượng khí khai thác tích lũy theo thời gian – LSTM	74
Hình 3-40: Dự báo sản lượng dầu khai thác theo mô hình GRU	75
Hình 3-41: Sản lượng dầu khai thác dự báo so với giá trị thực tế – LSTM	75
Hình 3-42: Sản lượng dầu khai thác tích lũy theo thời gian – GRU	76
Hình 3-43: Dự báo sản lượng khí khai thác theo mô hình GRU	76
Hình 3-44: Sản lượng khí khai thác dự báo so với giá trị thực tế – GRU	77
Hình 3-45: Sản lượng khí khai thác tích lũy theo thời gian – GRU	77

DANH MỤC HÌNH ẢNH PHỤ LỤC

Hình PL 1: Đồ thị semi – log của $\log(q)$ và t trong mô hình Exponential	88
Hình PL 2: Đồ thị $1/q$ với t trong mô hình Harmonic	88
Hình PL 3: Đồ thị q_t với t trong mô hình Hyperbolic	89
Hình PL 4: Phương trình $f(b)$ vô nghiệm trên khoảng $(0, 1)$	90
Hình PL 5: Đồ thị dự báo lưu lượng khai thác dầu theo thời gian trong phương pháp Exponential – DCA	92
Hình PL 6: Đồ thị dự báo lưu lượng khai thác dầu tích lũy theo thời gian trong phương pháp Exponential – DCA.....	93
Hình PL 7: Giá trị khai thác dầu dự báo so với giá trị khai thác dầu thực tế trong phương pháp Exponential - DCA	93
Hình PL 8: Đồ thị semi – log của $\log(q)$ và t trong mô hình Exponential	94
Hình PL 9: Đồ thị $1/q$ với t trong mô hình Harmonic	94
Hình PL 10: Đồ thị q_t với t trong mô hình Hyperbolic	95
Hình PL 11: Phương trình $f(b)$ vô nghiệm trên khoảng $(0, 1)$	96
Hình PL 12: Đồ thị dự báo lưu lượng khai thác khí theo thời gian trong phương pháp Exponential – DCA	98
Hình PL 13: Đồ thị dự báo lưu lượng khai thác khí tích lũy theo thời gian trong phương pháp Exponential – DCA.....	99
Hình PL 14: Giá trị khai thác khí dự báo so với giá trị khai thác khí thực tế trong phương pháp Exponential - DCA	99
Hình PL 15: Các ma trận huấn luyện mạng nơ – ron nhân tạo	100
Hình PL 16: Cửa sổ nn tool	101
Hình PL 17: Thiết kế cấu trúc mạng ANN.....	102
Hình PL 18: Cấu trúc mạng nơ – ron thiết kế	103
Hình PL 19: Quá trình huấn luyện mạng ANN	104

Hình PL 20: Biểu đồ sai số hội tụ của quá trình huấn luyện mạng trong dự báo sản lượng dầu khai thác	105
Hình PL 21: Biểu đồ hồi quy của các tập huấn luyện, xác thực và kiểm tra trong dự báo sản lượng dầu khai thác	105
Hình PL 22: Biểu đồ sai số hội tụ của quá trình huấn luyện mạng trong dự báo sản lượng khí khai thác.....	107
Hình PL 23: Biểu đồ hồi quy của các tập huấn luyện. xác thực và kiểm tra trong dự báo sản lượng khí khai thác	107

DANH MỤC BẢNG BIỂU

Bảng 1-1: Bảng tổng hợp các mô hình đường cong suy giảm thực nghiệm của Arps	8
Bảng 3-1: Thống kê đơn giản của bộ dữ liệu thô trước khi xử lý	44
Bảng 3-2: Thống kê đơn giản về dữ liệu thô sau khi xử lý	47
Bảng 3-3: Dữ liệu của giếng F15, mỏ Volve sau khi được tiền xử lý	50
Bảng 3-4: Bảng đánh giá mô hình hồi quy tuyến tính dự báo sản lượng dầu khai thác ...	60
Bảng 3-5: Bảng đánh giá mô hình hồi quy tuyến tính dự báo sản lượng khí khai thác	60
Bảng 3-6: Bảng đánh giá mô hình hồi quy Ridge dự báo sản lượng dầu khai thác	63
Bảng 3-7: Bảng đánh giá mô hình hồi quy Ridge dự báo sản lượng khí khai thác.....	63
Bảng 3-8: Bảng đánh giá mô hình mạng nơ – ron hồi quy đơn giản dự báo sản lượng dầu khai thác	67
Bảng 3-9: Bảng đánh giá mô hình mạng nơ – ron hồi quy đơn giản dự báo sản lượng khí khai thác.....	67
Bảng 3-10: Bảng đánh giá mô hình bộ nhớ ngắn hạn dài dự báo sản lượng dầu khai thác	71
Bảng 3-11: Bảng đánh giá mô hình bộ nhớ ngắn hạn dài dự báo sản lượng khí khai thác	71
Bảng 3-12: Bảng đánh giá mô hình nút hồi quy có cổng dự báo sản lượng dầu khai thác	74
Bảng 3-13: Bảng đánh giá mô hình nút hồi quy có cổng dự báo sản lượng khí khai thác	74
Bảng 3-14: Bảng đánh giá tổng hợp các mô hình học máy và học sâu trong dự báo sản lượng dầu khai thác	78
Bảng 3-15: Bảng đánh giá tổng hợp các mô hình học máy và học sâu trong dự báo sản lượng khí khai thác	78
Bảng 3-16: So sánh mô hình LSTM với phương pháp DCA và ANN trong dự báo sản lượng dầu khai thác	79
Bảng 3-17: So sánh mô hình LSTM với phương pháp DCA và ANN trong dự báo sản lượng khí khai thác	80

DANH MỤC BẢNG BIỂU PHỤ LỤC

Bảng PL 1: Kết quả dự báo khai thác dầu mỏ Volve với phương pháp DCA – mô hình Exponential	91
Bảng PL 2: Sai số đánh giá mô hình Exponential – phương pháp DCA trong dự báo khai thác dầu mỏ Volve	92
Bảng PL 3: Kết quả dự báo khai thác khí mỏ Volve với phương pháp DCA – mô hình Exponential	97
Bảng PL 4: Sai số đánh giá mô hình Exponential – phương pháp DCA trong dự báo khai thác khí mỏ Volve	98
Bảng PL 5: Đánh giá mô hình ANN trong dự báo khai thác dầu khí mỏ Volve	104
Bảng PL 6: Kết quả dự báo khai thác dầu mỏ Volve bằng phương pháp ANN	106
Bảng PL 7: Kết quả dự báo khai thác khí mỏ Volve bằng phương pháp ANN	108

DANH MỤC BIỂU ĐỒ

Biểu đồ 3-1: Biểu đồ boxplot và histogram của BTHP	45
Biểu đồ 3-2: Biểu đồ boxplot và histogram của CS	45
Biểu đồ 3-3: Biểu đồ boxplot và histogram của WHP	45
Biểu đồ 3-4: Biểu đồ boxplot và histogram của OIL	46
Biểu đồ 3-5: Biểu đồ boxplot và histogram của GAS	46
Biểu đồ 3-6: Biểu đồ boxplot và histogram của WAT	46

DANH MỤC KÝ HIỆU VIẾT TẮT VÀ THUẬT NGỮ

Thuật ngữ tiếng Anh	Ký hiệu	Thuật ngữ tiếng Việt
Adaptive Moment Estimation	Adam	(*)
Artificial Intelligence	AI	Trí tuệ nhân tạo
Artificial Neural Network	ANN	Mạng nơ – ron nhân tạo
Bottom Hole Flowing Pressure	BHFP	Áp suất dòng chảy đáy giếng
Bottom Hole Pressure	BTHP	Áp suất đáy giếng
Bottom Hole Temperature	BTHT	Nhiệt độ đáy giếng
Choke Size	CS	Độ mở của ống gốp
Decline Curve Analysis	DCA	Phân tích đường cong suy giảm
Deep Learning	DL	Học sâu
Different Pressure	DP	Chênh áp
Flowing Material Balance	FMB	Cân bằng vật chất dòng chảy
Gated Recurrent Unit	GRU	Nút hồi quy có cổng
Linear Regression	LR	Hồi quy tuyến tính
Long Short – Term Memory	LSTM	Bộ nhớ ngắn hạn dài
Mean Absolute Error	MAE	Sai số tuyệt đối trung bình
Machine Learning	ML	Học máy
Mean Squared Error	MSE	Sai số toàn phương trung bình
Correlation Coefficient	R	Hệ số tương quan
Coefficient of Determination	R ²	Hệ số xác định
Root Mean Squared Error	RMSE	Sai số toàn phương trung bình khai căn
Root mean square propagation	RMSprop	(**)
Recurrent Neural Network	RNN	Mạng nơ – ron hồi quy
Support Vector Machine	SVM	Máy véc – tơ hỗ trợ
Support Vector Regression	SVR	Hồi quy véc – tơ hỗ trợ
Wellhead Pressure	WHP	Áp suất đầu giếng
Wellhead Temperature	WHT	Nhiệt độ đầu giếng

(*), (**): Các phương pháp tối ưu trong học sâu

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Qua nhiều năm phát triển, trí tuệ nhân tạo dần được ứng dụng rộng rãi trong mọi lĩnh vực của đời sống. Không nằm ngoài xu hướng chung, việc ứng dụng trí tuệ nhân tạo vào lĩnh vực dầu khí cũng rất phổ biến những năm gần đây. Luận văn tập trung nghiên cứu về phương pháp phân tích khai thác dầu khí, trong đó ứng dụng công nghệ trí tuệ nhân tạo để nâng cao chất lượng dự báo khai thác.

Có một vấn đề được đặt ra cho các công ty dầu khí, rằng sản lượng khai thác trong tương lai của một mỏ có thể được dự báo trước hay không. Vấn đề này đã được tiếp cận qua nhiều phương pháp trong vài thập kỷ qua, tuy nhiên những phương pháp ấy thường như không phải là phương pháp tối ưu cho mọi mỏ dầu khí, mặt khác áp lực cắt giảm chi phí vận hành và tối đa hóa lợi nhuận của các công ty dầu khí cũng đã góp phần thúc đẩy các nhà nghiên cứu tìm ra những phương pháp dự báo thay thế.

Các phương pháp dự báo sản lượng khai thác truyền thống chưa thực sự đưa ra dự báo sát với thực tế. Để tăng hiệu quả dự báo, đồng thời có nhiều cơ sở đổi sánh và nhiều phương pháp thay thế khác nhau, các phương pháp dự báo dựa trên học máy nói chung cũng được tích cực phát triển.

Việc dự báo sản lượng khai thác dầu khí của những mỏ gần hết vòng đời mỏ cũng rất quan trọng. Từ đó người ta có thể suy ra ngày đóng giếng hoặc chuyển đổi mục đích sử dụng giếng, chuyển đổi chế độ khai thác một cách hợp lý.

2. Mục tiêu nghiên cứu

Xây dựng các mô hình dự báo sản lượng khai thác dầu khí cho giếng F14, mỏ Volve, biển Bắc, Na Uy dựa trên tập dữ liệu chứa các thông số về áp suất, nhiệt độ, lưu lượng khai thác theo thời gian của giếng trong suốt quá trình khai thác. Các mô hình này được xây dựng dựa trên các phương pháp học máy hồi quy, bao gồm hồi quy tuyến tính, máy véc – tơ hỗ trợ; và mạng nơ – ron hồi quy, bao gồm: mạng nơ – ron hồi quy đơn giản (Simple RNN), bộ nhớ ngắn hạn dài (LSTM) và nút hồi quy có cổng (GRU).

3. Nhiệm vụ của luận văn

- Trình bày lịch sử phát triển và cơ sở lý thuyết các phương pháp dự báo sản lượng khai thác truyền thống.
- Trình bày cơ sở lý thuyết của các phương pháp học máy và học sâu nêu trên.
- Ứng dụng các phương pháp nêu trên, xây dựng mô hình dự báo bằng ngôn ngữ lập trình Python.
- Kiểm nghiệm và đánh giá mức độ hiệu quả của các mô hình.

4. Phương pháp nghiên cứu

- Trong các phương pháp học máy và học sâu, chọn ra những phương pháp tiềm năng cho việc nghiên cứu. Các mô hình được chọn ở đây là: hồi quy tuyến tính, máy véc – tơ hỗ trợ, mạng nơ – ron hồi quy đơn giản (Simple RNN), bộ nhớ ngắn hạn dài (LSTM) và nút hồi quy có công (GRU).
- Xây dựng mô hình dự báo dựa trên các phương pháp nêu trên theo nguyên tắc “thử và sai”. Việc “thử và sai” có thể được thực hiện ở hầu hết các khâu: tiền xử lý dữ liệu, đặc tả mô hình, xác thực mô hình, kiểm tra mô hình.
- Quá trình thử và sai chỉ dừng lại khi ta nhận được những kết quả nằm trong giới hạn cho phép của những tiêu chí đặt ra.

5. Ý nghĩa khoa học và thực tiễn

Về khoa học, luận văn đã thành công ứng dụng các phương pháp dự báo sản lượng dầu khí khai thác. Trong đó, những ưu điểm và nhược điểm của từng phương pháp đã xem xét và đánh giá cụ thể, từ đó đưa ra phương pháp dự báo tốt nhất. Bên cạnh đó, nó cũng làm đầy đủ hơn và đa dạng hơn tập hợp những phương pháp cùng giải quyết một vấn đề chung, đó là dự báo sản lượng khai thác dầu khí. Việc ứng dụng thành công các phương pháp học máy và học sâu trong dự báo sản lượng dầu khí khai thác đã làm tăng độ tin cậy của phương pháp này

Về thực tiễn, luận văn đã đưa ra những phương pháp dự báo hiệu quả đối với giếng F14, mỏ Volve. Bằng các phương pháp thống kê, luận văn đã chỉ ra những đặc điểm tương quan của những đặc trưng liên quan đến quá trình khai thác trong bộ dữ liệu đầu vào, từ đó

chọn lọc những đặc trưng có ảnh hưởng lẫn nhau để đưa vào các mô hình học máy, học sâu. Đồng thời, nó cũng mở ra triển vọng dự báo khai thác đối với những giếng khác của mỏ Volve và các mỏ lân cận với cùng phương pháp nêu trên.

6. Tổng quan tình hình nghiên cứu

6.1. R. G. Agatwal, D. C. Gardner, and S. W. Kleinstreiber, Analyzing Well production data using combined type curve and decline curve analysis concepts, 1998 [1]

Cùng với sự phát triển độc lập của hai phương pháp dự báo sản lượng khai thác dầu khí truyền thống là phân tích đường cong suy giảm (DCA) và phân tích đường cong loại (Type curves), năm 1996 Agatwal, Gardner và Kleinstreiber đã phát triển một phương pháp kết hợp của hai phương pháp nói trên. Bài báo trình bày các đường cong suy giảm sản lượng để phân tích dữ liệu khai thác từ các giếng dầu khí của các vỉa đứt gãy.

Nghiên cứu có những cải tiến đáng kể so với hai phương pháp truyền thống riêng lẻ. Một tập hợp mới các đường cong loại suy giảm sản lượng theo thời gian – lưu lượng, thời gian tích lũy – lưu lượng tích lũy và các dẫn xuất liên quan của chúng đã được phát triển bằng cách sử dụng các khái niệm phân tích chuyển tiếp áp suất (pressure transient analysis). Ngoài ra, các đường cong kiểu suy giảm sản lượng này có thể phân biệt rõ ràng hơn giữa các giai đoạn dòng chảy chịu sự ảnh hưởng của điều kiện chuyển tiếp và điều kiện biên (transient and boundary dominated flow periods). Chúng cung cấp một công cụ thực tế cho các kỹ sư mỏ để ước tính dầu khí tại chỗ, cũng như ước tính độ thấm của vỉa, hiệu ứng skin, chiều dài đứt gãy và độ dẫn đứt gãy.

6.2. R. J. Boomer and T. Exploration, Predicting Production Using a Neural Network (Artificial Intelligence Beats Human Intelligence) Society of Petroleum Engineers, 1995 [2]

Năm 1995, Boomer lần đầu tiên đề xuất phương pháp dự báo sản lượng dầu khí khai thác với mạng nơ – ron nhân tạo. Các phương pháp dự báo của chuyên gia ở thời điểm đó có nhiều sai số so với thực tế như đánh giá quá thấp hoặc quá cao tiềm năng thực sự của một giếng, một mỏ dầu khí, khiến cho việc khai thác không đạt được lợi ích tối đa, thậm chí không thể hòa vốn. Trước nhu cầu cần kíp là phải tìm ra được phương pháp dự báo tối ưu cho công tác dự báo khai thác, Boomer đã đề xuất mạng nơ – ron nhân tạo, tạo tiền đề cho phương pháp học máy và học sâu phát triển.

Trong nghiên cứu này, ông đã đưa ra khái niệm “mặt nạ dữ liệu” (data mask). Mặt nạ dữ liệu là một lưới 5×5 bao phủ 1.000 mẫu Anh. Mỗi ô vuông (1.320 feet mỗi cạnh) tương đương với 40 mẫu Anh. Mặt nạ dữ liệu này được đặt trên khu vực rộng 40 mẫu Anh của vị trí khoan, được thu thập cho mọi giếng trong ranh giới của nó. Ông chia lưới trên thành 3 “vòng đồng tâm”, nghĩa là 5 ô lưới ngoài cùng trên mỗi cạnh tạo thành vòng ngoài, 3 ô tiếp theo tạo thành vòng giữa và ô cuối cùng ở tâm chứa vị trí của giếng cần được dự báo. Dữ liệu thu thập được từ các giếng của mỗi vòng được tính trung bình và đưa vào mô hình mạng nơ – ron nhân tạo để huấn luyện. Đầu ra của mạng nơ – ron nhân tạo là sản lượng khai thác tích lũy trong một, ba, sáu và mười hai tháng cho vị trí khoan được đề xuất.

Với mỏ Vacuum, phương pháp mạng nơ – ron nhân tạo cho sai số nhỏ hơn 3.5 lần so với phương pháp của chuyên gia và đạt đến 93% hiệu quả dự báo.

6.3. J. Sun, X. Ma, and M. Kazi, Comparison of Decline Curve Analysis DCA with Recursive Neural Networks RNN for Production Forecast of Multiple Wells, 2018 [3]

Tác giả đã ứng dụng một nhánh của mạng nơ – ron nhân tạo là bộ nhớ ngắn hạn dài (LSTM) thuộc nhóm mạng nơ – ron hồi quy (RNN) để xây dựng mô hình dự báo sản lượng của các sản phẩm khai thác. Bên cạnh đó, tác giả cũng đã phân tích khai thác với các phương pháp phân tích đường cong suy giảm cải tiến như Duong, SEPD và PLE làm cơ sở so sánh.

Tập dữ liệu khai thác được sử dụng trong bài báo là từ Eagle Ford Shale play ở tây Texas với khoảng 600 – 800 điểm dữ liệu lịch sử theo ngày. Dữ liệu khai thác dầu, khí và nước hàng ngày đều được ghi lại bao gồm áp suất đầu giếng. Các dữ liệu khai thác này là các biến đầu vào chính cho mô hình mạng nơ – ron. Mô hình cơ sở được xây dựng cho việc dự báo một giếng duy nhất. Hơn nữa, kịch bản giếng tổng hợp cũng được xem xét để cải thiện mô hình với một số dữ liệu lịch sử khai thác của giếng lân cận cũng được được xem như là biến đầu vào. Các đầu ra giống nhau được chọn để kiểm tra xem các mô hình đã phát triển có cho ra kết quả nhất quán và hiệu quả hay không.

Cách tiếp cận bằng mạng nơ – ron nhân tạo này đã được xác thực và so sánh các phương pháp DCA truyền thống, cho thấy khả năng dự báo xu hướng tốt hơn nhiều và tính toán có sai số nhỏ hơn. Nó cũng xem xét các tình huống vận hành phức tạp có thể là đặc điểm của hò chúa khác thường mà các phương pháp DCA không thể thực hiện. So với mô hình dựa trên vật lý mô phỏng số học, phương pháp ANN không phụ thuộc nhiều vào các tính toán dựa nguyên lý kỹ thuật và đưa ra các giải pháp nhanh hơn. Mặc dù cách tiếp cận theo hướng dữ liệu này không có tính vật lý, nhưng nó là một cách thay thế tốt để cung cấp các kết quả nhanh chóng và mạnh mẽ bên cạnh các mô phỏng phản ánh số và thực nghiệm.

6.4. Y. Li, R. Sun, and R. Horne, Deep learning for well data history analysis, 2019 [4]

Trong nghiên cứu này, tác giả đã phát triển các quy trình xử lý dữ liệu cho dữ liệu kiểu chuỗi thời gian và xác định việc lựa chọn tần suất lấy mẫu dữ liệu, tổ hợp tham số và cấu trúc dữ liệu phù hợp cho các mô hình học sâu. Sau đó, họ ứng dụng các mô hình học sâu gồm hai mô hình mạng nơ – ron hồi quy (RNN) là bộ nhớ ngắn hạn dài LSTM và nút hồi quy có cổng (GRU), để phân tích dữ liệu giềng (áp suất) và kết hợp các mô hình vật lý và các mô hình học sâu ấy. Cách tiếp cận này bảo toàn thông tin trước đó và tạo ra phản ứng hiện tại với bộ nhớ về ứng xử của giềng trước đó. Ngoài ra, một sự kết hợp mới lạ giữa RNN với Mạng nơ – ron tích chập (CNN) được gọi là mạng chuỗi thời gian dài và ngắn hạn (LSTNet), cũng đã được khảo sát.

Cả GRU và LSTM đều chính xác hơn một cấu trúc RNN đơn giản. GRU nhanh hơn một chút so với LSTM vì số lượng cổng ít hơn trong LSTM. Đối với các trường hợp mở phức tạp hơn, LSTNet đã chứng minh các cải tiến hiệu suất đáng kể so với các phương pháp cơ sở.

6.5. H. Alimohammadi, H. Rahmanifard, and N. Chen, Multivariate time series modelling approach for production forecasting in unconventional resources, 2020 [5]

Tác giả đã ứng dụng hai phương pháp học sâu thuộc họ mạng nơ – ron hồi quy là mạng nơ – ron hồi quy hai chiều (BRNN), bộ nhớ ngắn hạn dài (LSTM) và nút hồi quy có cổng (GRU) để xây dựng mô hình dự báo sản lượng khai thác.

Ba mô hình học sâu được huấn luyện trên tập dữ liệu thu thập từ một giềng ngang nhiều vết nứt với các thông số: lưu lượng khai thác đa pha trong một năm, nhiệt độ dòng chảy và áp suất đường ống. Bộ dữ liệu cho thấy lưu lượng khai thác biến động mạnh sau 250 ngày khai thác, điều này làm cho xu hướng suy giảm lưu lượng khó được nắm bắt bằng các phương pháp DCA thông thường.

Phương pháp này có thể được sử dụng để dự báo lưu lượng khai thác trong tương lai và cũng có thể được sử dụng để ước lượng dữ liệu lịch sử lưu lượng bị thiếu. Dự báo lưu lượng sử dụng mô hình học sâu linh động hơn và có cơ hội tốt hơn để nắm bắt các sự kiện phi tuyến với khả năng xử lý lượng dữ liệu động không lồ. Quy trình công việc đề xuất đã được áp dụng thành công để dự đoán dòng chảy ba pha chỉ dựa trên hai đầu vào động. Khả năng dự đoán sớm của mô hình được chứng minh bằng cách dự đoán quá trình sản xuất dài hạn khi nó được huấn luyện với dữ liệu khai thác một và hai tháng.

7. Cấu trúc luận văn

Cấu trúc luận văn bao gồm 3 chương:

Chương 1 – TỔNG QUAN VỀ DỰ BÁO KHAI THÁC DẦU KHÍ: Khái quát lại lịch sử phát triển của các phương pháp dự báo sản lượng dầu khí trong khai thác và nêu lại 2 phương pháp dự báo truyền thống là Đường cong suy giảm.

Chương 2 – TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO: Trình bày cơ sở lý thuyết về các phương pháp học máy và học sâu nêu trên.

Chương 3 – PHƯƠNG PHÁP THỰC HIỆN: Trình bày tổng quan về mô hình Volve, quy trình xây dựng mô hình học máy, học sâu và đánh giá các kết quả đạt được.

Ngoài ra, như đã đề cập ở trên, luận văn cũng tiến hành khảo sát phương pháp đường cong suy giảm và xây dựng mô hình mạng nơ – ron nhân tạo bằng matlab để làm cơ sở so sánh cho các phương pháp khảo sát chính.

CHƯƠNG 1: TỔNG QUAN VỀ DỰ BÁO KHAI THÁC DẦU KHÍ

1.1 Lịch sử phát triển các phương pháp phân tích quá trình khai thác

Việc phân tích khai thác bắt đầu vào năm 1920 với mục tiêu đơn thuần là tìm ra phương trình suy giảm tốt nhất để có thể dự đoán doanh thu của hoạt động khai thác trong tương lai trên cơ sở thực nghiệm mà không dựa trên một nền tảng kỹ thuật nào. Sau đó, vào năm 1945, Arps đã xây dựng phương trình suy giảm theo hàm mũ, hyperbol và harmonic với áp suất không đổi. Năm 1960, Fetkovich giới thiệu phương pháp type curves và cũng với giả thiết áp suất dòng chảy không đổi theo thời gian. Năm 1993, Palacio và Blasingame đã giới thiệu một phương pháp type curves với lưu lượng và áp suất thay đổi dưới dạng đồ thị log – log của chỉ số khai thác và thời gian cân bằng vật chất.

Hiện nay, kỹ thuật phân tích sự suy giảm khai thác bao gồm phương pháp Arps truyền thống (1945), phương pháp đổi sánh type curves cổ điển của Fetkovich (1980), phương pháp đổi sánh type curves hiện đại của Palacio và Blasingame (1993) với Agarwal (1998) và phương pháp kỹ thuật via Flowing Material Balance – FMB (1998).

Việc ngoại suy xu hướng của một số biến của giếng là một phương pháp vô cùng hữu ích cho việc phân tích dữ liệu khai thác khi ta không có nhiều thông tin về vỉa. Trong việc phân tích dữ liệu khai thác, biến đơn giản nhất và có giá trị sử dụng cao nhất là sản lượng khai thác của nó. Các xu hướng hoặc quan hệ toán học có thể được đúc kết thông qua lịch sử khai thác của một giếng bất kỳ, sử dụng để dự báo hiệu suất khai thác trong tương lai, được gọi là phương pháp phân tích đường cong suy giảm Arps truyền thống, hay còn gọi là DCA. Phương pháp này cho biết quy luật suy giảm sản lượng của giếng với áp suất dòng chảy ở đáy giếng (BHFP) không đổi và dòng chảy ở trong giai đoạn chịu ảnh hưởng biên (boundary - dominated flow period). Ưu điểm của phương pháp này là ta không nhất thiết phải thu thập các dữ liệu về thành hệ mới có thể xây dựng mô hình dự báo. Bên cạnh đó, phương pháp này không thích hợp để phân tích dữ liệu trong giai đoạn dòng chuyển tiếp.

Năm 1980, Fetkovich đã đưa công thức dòng chảy chuyển tiếp trong phân tích well test vào phân tích đường cong suy giảm, do đó phương pháp type curves của Arps được mở rộng đến giai đoạn dòng chảy chuyển tiếp (transient flow period). Bằng cách này, quy luật suy giảm sản lượng khai thác và ảnh hưởng biên được thể hiện trực quan. Ưu điểm lớn nhất của phương pháp này là khả năng xác định đáng tin cậy việc khai thác đang ở trong giai đoạn dòng chảy chuyển tiếp hay trong giai đoạn dòng chảy chịu ảnh hưởng biên.

Cả hai phương pháp Arps và Fetkovich đều giả định rằng BHFP là không đổi cho việc phân tích dữ liệu khai thác mà không tính đến sự thay đổi các đặc điểm PVT của khí so với áp suất đáy giếng. Năm 1993, Palacio và Blasingame đã giới thiệu pseudo - pressure normalized production ($q/\Delta p_p$) và material balance pseudo-time (t_{CA}) để xây dựng type-curves, trong đó xem xét đến việc khai thác ở nhiều BHFP khác nhau và PVT khí thay đổi theo áp suất thành hệ.

Năm 1998, Agarwal và cộng sự đã sử dụng các quan hệ của pseudo - pressure normalized production ($q/\Delta p_p$), material balance pseudo - time (t_{CA}), và các thông số không thứ nguyên trong việc phân tích well test để xây dựng phương pháp phân tích suy giảm sản lượng Agarwal - Gardner. Cả hai phương pháp Blasingame và Agarwal - Gardner đều sử dụng pseudo - pressure normalized production ($q/\Delta p_p$) và material balance pseudo - time (t_{CA}) để xây dựng đường cong, trong khi phương pháp NPI/Normalized Pressure Integral (Blasingame và cộng sự, 1989) sử dụng tích phân $q/\Delta p_p$ để phân tích dữ liệu có sẵn, không bị ảnh hưởng bởi sự phân tán dữ liệu.

Phương pháp phân tích đối sánh type curve của Palacio và Blasingame (1993) và Agarwal (1998) đã sử dụng material balance pseudo - time và pseudo - pressure normalized production để giải quyết vấn đề thay đổi của BHFP, lưu lượng và PVT của khí. Họ đã sử dụng tích phân lưu lượng dòng chảy, đạo hàm tích phân lưu lượng dòng chảy, thời gian khai thác tích lũy và đường cong của sản lượng khai thác tích lũy với lưu lượng dòng chảy làm đường cong phân tích đối sánh phụ trợ để nhằm làm sáng tỏ hơn cách diễn giải kết quả.

Mattar (1998, 2006) và Agarwal (1998) đề xuất sử dụng phương pháp “flow (dynamic) material balance” để phân tích dữ liệu khai thác, sau đó tiến hành phân tích chi tiết về việc tính toán thời gian cân bằng vật chất. Phương pháp này khá đơn giản và dễ tiếp cận. Mattar và Anderson (2003) tin rằng không có một phương pháp phân tích dữ liệu khai thác chung nào có thể đáp ứng tất cả các loại vía, và cách tốt nhất để loại bỏ sai số phân tích là sử dụng tổng hợp tất cả các phương pháp phân tích, kết hợp xem xét dữ liệu áp suất dòng chảy.

Trải qua gần một thế kỷ, kỹ thuật phân tích khai thác đã phát triển với một số tiến bộ, bao gồm cả mục tiêu được phân tích, tức là, từ dữ liệu khai thác thuần túy sang cả dữ liệu lưu lượng dòng chảy và áp suất; mô hình phân tích, nghĩa là từ không có mô hình sang có cả mô hình phân tích và mô hình số; phương pháp phân tích, nghĩa là, từ phương pháp Arps thực nghiệm đến phương pháp log – log được mà đại diện là Blasingame; các điều kiện áp dụng, nghĩa là, từ dữ liệu khai thác áp suất không đổi đơn giản đến dữ liệu áp suất thay đổi và lưu lượng thay đổi; và các thông số ước tính, nghĩa là, từ chỉ khai thác tích lũy đến nhiều thông số như độ thẩm thành hệ, yếu tố skin, trữ lượng động (dynamic reserves), khu vực tháo khô (drainage area), cũng như các yếu tố liên kết giữa các giếng và tiềm năng bơm ép.

1.2 Phân tích đường cong suy giảm Arps [6]

Đối với các giếng có lịch sử khai thác lâu dài và khai thác trong điều kiện áp suất đáy giếng (BHFP) không đổi, năm 1945 Arps đưa ra ba loại suy giảm sản lượng theo mối quan hệ lưu lượng - thời gian, bao gồm suy giảm theo mô hình exponential, hyperbol và harmonic. Phương pháp này đơn giản, không cần quan tâm đến các thông số của vỉa hoặc giếng và có thể áp dụng cho các loại vỉa khác nhau. Tuy nhiên, phương pháp này có ba hạn chế:

- Đầu tiên, lượng dự trữ tối đa có thể thu hồi phải được tính toán với giả định rằng các điều kiện khai thác sẽ không thay đổi trong tương lai.
- Thứ hai, đồ thị đường cong suy giảm cho biết quy luật suy giảm khi dòng chảy ở trạng thái chịu ảnh hưởng biên; do đó, nó không thể được sử dụng để phân tích dữ liệu khi mà dòng chảy đang trong giai đoạn chuyển tiếp.
- Áp suất đáy giếng không đổi theo thời gian

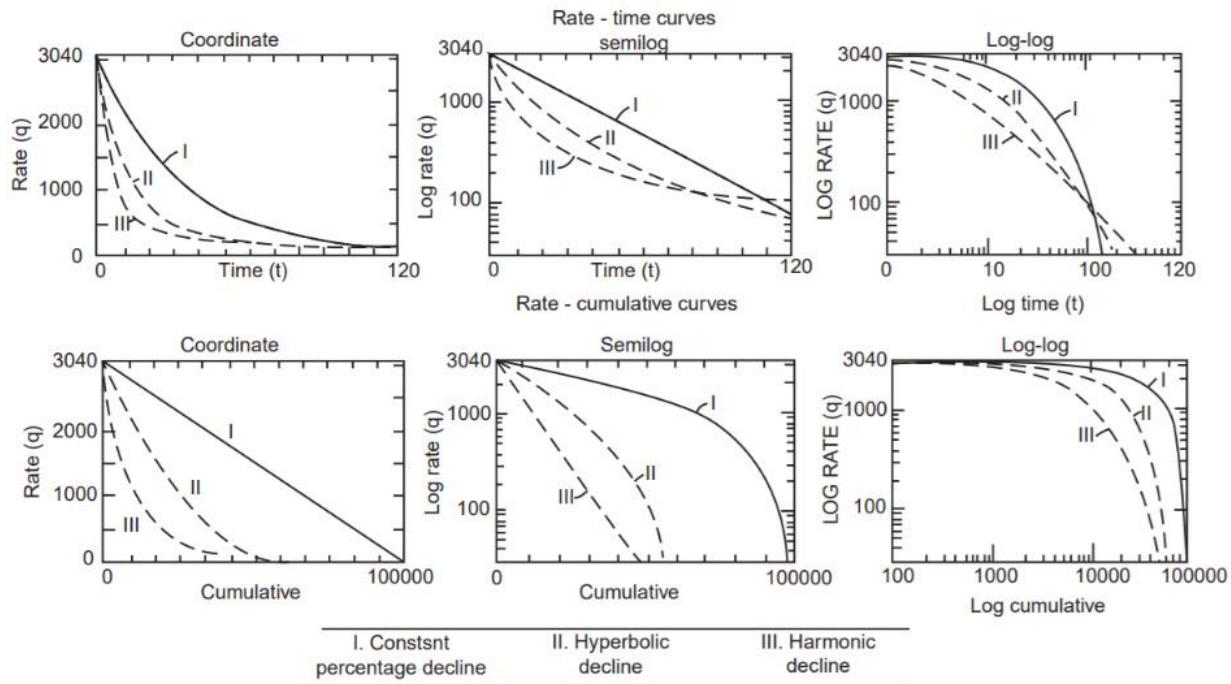
Phương pháp ngoại suy xu hướng nhằm mục đích ước tính khả năng khai thác trong tương lai phải thỏa mãn các yếu tố gây ra sự thay đổi trong hoạt động khai thác trong quá khứ, ví dụ, sự suy giảm lưu lượng dòng chảy, sẽ hoạt động theo cùng một cách trong tương lai. Các đường cong suy giảm này được đặc trưng bởi ba yếu tố:

- Lưu lượng khai thác ban đầu hoặc lưu lượng khai thác tại một số thời điểm cụ thể
- Độ cong
- Lưu lượng suy giảm

Arps (1945) đề xuất rằng độ cong trong đường cong lưu lượng khai thác so với thời gian có thể được biểu thị bằng họ phương trình họ hyperbol. Arps nhận ra có ba loại suy giảm lưu lượng sau: Exponential, Harmonic, Hyperbolic.

Mỗi loại đường cong suy giảm có một độ cong khác nhau, như trong hình 1.1. Hình này mô tả hình dạng đặc trưng của từng loại suy giảm khi lưu lượng dòng chảy được lập biểu đồ theo thời gian hoặc so với sản lượng tích lũy trên các thang đo Descartes, semilog và log – log. Các đặc điểm chính của các đường cong suy giảm này có thể được sử dụng để chọn mô hình suy giảm lưu lượng dòng chảy phù hợp để mô tả mối quan hệ lưu lượng – thời gian hoặc lưu lượng – sản lượng khai thác tích lũy:

- Exponential decline: Mối quan hệ tuyến tính được biểu diễn khi lưu lượng dòng chảy so với thời gian được vẽ trên thang đo semilog và cũng được biểu diễn trong trường hợp lưu lượng dòng chảy so với sản lượng tích lũy được vẽ trên thang đo Descartes.
- Harmonic decline: Lưu lượng so với sản lượng tích lũy là một đường thẳng trên thang đo semilog. Tất cả các loại đường cong suy giảm khác đều có một số độ cong nhất định. Do đó có nhiều kỹ thuật chuyển đổi được thiết kế nhằm “làm thẳng” đường cong, nó là kết quả của việc biểu diễn đồ thị lưu lượng dòng chảy so với thời gian trên thang log – log.
- Hyperbolic decline: Không có thang đo biểu đồ nào ở trên (Descartes, semilog hoặc log – log) biểu diễn được mối quan hệ tuyến tính cho sự suy giảm hyperbol. Tuy nhiên, nếu lưu lượng dòng chảy được biểu diễn theo thời gian trên đồ thị log – log, đường cong kết quả có thể được “làm thẳng” bằng các kỹ thuật chuyển đổi.

**Hình 1-1:** Các loại đường cong suy giảm

Phân tích đường cong suy giảm dựa trên các mối quan hệ thực nghiệm của lưu lượng khai thác so với thời gian, được Arps (1945) đưa ra như sau:

$$q_t = \frac{q_i}{(1+bD_i t)^{1/b}} \quad (1-1)$$

Trong đó: $q(t)$: lưu lượng ở thời điểm t ($m^3/ngày$)

q_i : lưu lượng ban đầu ($m^3/ngày$)

D_i : Tốc độ suy giảm ban đầu ($1/ngày$)

b : hệ số mũ của đường cong suy giảm Arps

t : thời gian (ngày)

Tốc độ suy giảm D được định nghĩa là tỉ lệ của sự thay đổi logarithm tự nhiên của lưu lượng khai thác, tức là, $\ln(q)$, theo thời gian, t:

$$D = -\frac{d(\ln q)}{dt} = -\frac{1}{q} \frac{dq}{dt} \quad (1-2)$$

$$b = \frac{d}{dt} \left(\frac{1}{D} \right) = -\frac{d}{dt} \left(\frac{q dt}{dq} \right) \quad (1-3)$$

Dấu trừ đã được thêm vào vì dq và dt trái dấu, mặt khác ta có D luôn luôn dương. Phương trình tốc độ suy giảm 1.2 cho biết những thay đổi tức thời của hệ số góc của đường cong dq/dt , so với sự thay đổi của tốc độ dòng chảy q theo thời gian.

Phương pháp phân tích đường cong suy giảm này có thể được áp dụng cho các giếng riêng lẻ hoặc toàn bộ vỉa. Dựa trên ứng xử suy giảm lưu lượng của vỉa, giá trị của b nằm trong khoảng từ 0 đến 1, và do đó, phương trình Arps có thể được biểu diễn thuận tiện dưới ba dạng sau:

Bảng 1-1: Bảng tổng hợp các mô hình đường cong suy giảm thực nghiệm của Arps

Mô hình	Exponential	Hyperbolic	Harmonic
Hệ số đặc trưng	D là hằng số $b = 0$ $D_i = \frac{1}{(t_2 - t_1)} \ln\left(\frac{q_1}{q_2}\right)$	D phụ thuộc b $0 < b < 1$ $D_i = \frac{\left(\frac{q_i}{q_t}\right)^b - 1}{dt}$	D tỷ lệ với lưu lượng $b = 1$ $D_i = \frac{q_1 - 1}{q_2}$
Hàm lưu lượng theo thời gian q_t	$q = q_i e^{-D_i t}$	$q = \frac{q_i}{(1 + b D_i t)^{\frac{1}{b}}}$	$q = \frac{q_i}{1 + D_i t}$
Hàm sản lượng khai thác tích lũy theo lưu lượng	$Q = \frac{q_i - q_{ab}}{D_i}$	$Q = \frac{q_i}{(1 - b) D_i} \left[1 - \left(\frac{q_t}{q_i} \right)^{1-b} \right]$	$Q = \frac{q_i}{D_i} \ln\left(\frac{q_i}{q_t}\right)$
Ứng dụng mô hình	Xác định trữ lượng tối thiểu	Xác định trữ lượng tiềm năng	Xác định trữ lượng có thể thu hồi được tại thời điểm đang xét
Lượng Hydrocacbon có thể thu hồi	$N_p + \frac{q_i - q_{ab}}{D_i}$	$N_p + \frac{q_i}{(1 - b) D_i} \left[1 - \left(\frac{q_t}{q_i} \right)^{1-b} \right]$	$N_p + \frac{q_i}{D_i} \ln\left(\frac{q_i}{q_t}\right)$

Trong đó, q_{ab} là lưu tại thời điểm hủy giếng t_{ab} .

Có thể chỉ ra rằng ba dạng phương trình đường cong suy giảm này chỉ có thể áp dụng khi giếng/vỉa trong điều kiện dòng chảy trạng thái giả ổn định (bán ổn định), nghĩa là, điều kiện dòng chảy chịu ảnh hưởng biên (boundary - dominated flow conditions). Sau đây là danh sách các giả định cần phải được thỏa mãn trước khi thực hiện phân tích đường cong lưu lượng - thời gian suy giảm:

- Giếng tháo khô một khu vực tháo khô liên tục (drainage area), tức là giếng ở trong tình trạng dòng chảy chịu ảnh hưởng biên.
- Giếng được khai thác ở mức công suất tối đa hoặc gần tối đa.
- Giếng được khai thác với áp suất đáy giếng không đổi.

Ba điều kiện này phải được thỏa mãn trước khi áp dụng bất kỳ phương pháp phân tích đường cong suy giảm nào để mô tả hiệu suất khai thác của vỉa. Tuy nhiên, có thể rất khó xác định khi nào giếng đã xác định được khu vực tháo khô của nó để xác định điều kiện dòng chảy ở trạng thái giả ổn định. Đó là trở ngại chính của việc sử dụng đường cong suy giảm của Arps.

CHƯƠNG 2: TỔNG QUAN VỀ TRÍ TUỆ NHÂN TẠO

2.1 Giới thiệu chung

Trí tuệ nhân tạo là một nhánh của khoa học máy tính nhằm mục đích tạo ra những cỗ máy thông minh. Nó đã trở thành một phần không thể thiếu của các ngành công nghệ hiện đại. Trí tuệ nhân tạo (AI) là sự mô phỏng các quá trình thông minh của con người bằng máy móc, đặc biệt là hệ thống máy tính. Các quá trình này bao gồm học tập (thu nhận và xử lý thông tin), suy luận (sử dụng các quy tắc tổng quát để đưa ra kết luận gần đúng hoặc xác định) và tự điều chỉnh.

Học máy (machine learning) là một lĩnh vực con của trí tuệ nhân tạo. Các thuật toán học máy xây dựng một mô hình dựa trên dữ liệu mẫu, được gọi là tập dữ liệu huấn luyện, để đưa ra dự đoán hoặc quyết định mà không cần phải lập trình rõ ràng. Học máy được phân thành nhiều loại khác nhau dựa trên các phương pháp tiếp cận sau:

Học có giám sát: Các thuật toán học có giám sát xây dựng mô hình toán học của một tập dữ liệu chứa cả đầu vào và đầu ra mong muốn. Dữ liệu huấn luyện bao gồm một tập hợp các ví dụ huấn luyện. Mỗi ví dụ huấn luyện có một hoặc nhiều đầu vào và đầu ra mong muốn, còn được gọi là tín hiệu giám sát. Thông qua quá trình tối ưu hóa hàm mục tiêu, các thuật toán học có giám sát có thể được sử dụng để dự đoán đầu ra dựa trên một đầu vào mới.

Học không giám sát: Các thuật toán học tập không giám sát nhận một tập dữ liệu chứa đầu vào chưa được gán nhãn, từ đó suy ra cấu trúc trong tập dữ liệu, như nhóm hoặc phân cụm các điểm dữ liệu. Các thuật toán học tập không giám sát chỉ đơn thuần xác định các đặc điểm tương đồng trong tập dữ liệu ban đầu và phản ứng dựa trên sự hiện diện hoặc vắng mặt của các đặc điểm đó đối với mỗi điểm dữ liệu mới.

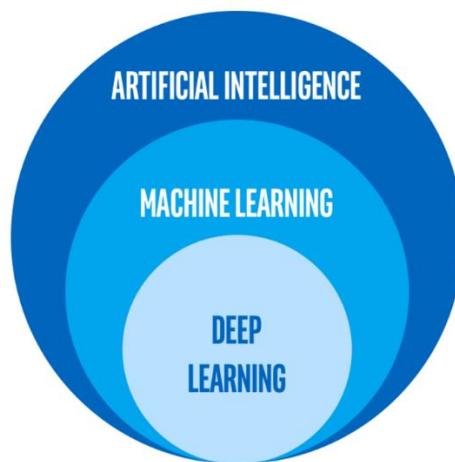
Học bán giám sát: Học bán giám sát là phương pháp học trung gian của học không giám sát và học có giám sát. Trong học bán giám sát, một số ví dụ huấn luyện bị thiếu nhãn. Nhiều nhà nghiên cứu học máy đã chỉ ra rằng dữ liệu không được gắn nhãn, khi được

sử dụng cùng với một lượng nhỏ dữ liệu được gắn nhãn, có thể tạo ra sự cải thiện đáng kể về độ chính xác của việc học.

Học tăng cường: Học tăng cường nghiên cứu cách thức một thực thể trong một môi trường nên chọn thực hiện các hành động nào để cực đại hóa một khoản thưởng nào đó về lâu dài. Các thuật toán học tăng cường cố gắng tìm một chiến lược ánh xạ các trạng thái của môi trường tới các hành động mà thực thể nên chọn trong các trạng thái đó.

Trong học máy, học đặc trưng (feature learning) hoặc học đại diện (representation learning) là một tập hợp các kỹ thuật cho phép hệ thống tự động khám phá các đại diện cần thiết để phát hiện hoặc phân loại đặc trưng từ dữ liệu thô. Điều này thay thế kỹ thuật đặc trưng (feature engineering) thủ công và cho phép một máy vừa học các đặc trưng vừa sử dụng chúng để thực hiện một tác vụ cụ thể.

Cùng với học đặc trưng, học sâu (deep learning) cũng là một nhánh của học máy. Học sâu được xây dựng dựa trên mạng nơ – ron nhân tạo với học đặc trưng. Việc học có thể được giám sát, bán giám sát hoặc không giám sát. Học sâu có nhiều kiến trúc khác nhau như mạng nơ – ron sâu (deep neural networks), mã mạng nơ – ron tích chập (convolutional neural networks), mạng niềm tin sâu (deep belief networks) và mạng nơ – ron hồi quy (recurrent neural networks).

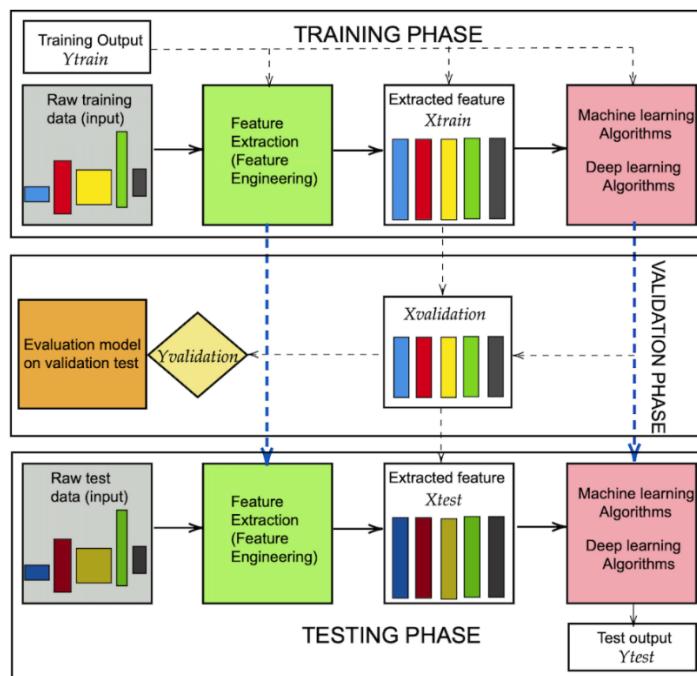


Hình 2-1: Mối quan hệ giữa trí tuệ nhân tạo, học máy và học sâu

2.2 Kỹ thuật xây dựng đặc trưng [7]

Mỗi điểm dữ liệu trong một mô hình học máy thường được biểu diễn bằng một vector gọi là vector đặc trưng. Trong cùng một mô hình các vector đặc trưng của các điểm thường có cùng kích thước. Điều này rất quan trọng vì các mô hình học máy bao gồm nhiều phép toán trên ma trận, các phép toán này yêu cầu tập dữ liệu có chiều phù hợp. Tuy nhiên, dữ liệu thực tế thường ở dạng thô với kích thước khác nhau nhưng có cùng chiều nhưng kích thước quá lớn gây trở ngại cho việc lưu trữ. Vì vậy, việc lựa chọn, tính toán đặc trưng phù hợp cho mỗi bài toán là một bước quan trọng.

Phần lớn các mô hình học máy có thể được minh họa như trong hình 3.2. Có hai pha lớn trong mỗi bài toán học máy là pha huấn luyện (training phase) và pha kiểm tra (testing phase). Pha huấn luyện xây dựng mô hình dựa trên dữ liệu huấn luyện. Dữ liệu kiểm tra được dùng cho việc đánh giá mô hình.



Hình 2-2: Mô hình học máy nói chung

Các điểm dữ liệu đôi khi được đo bằng những đơn vị khác nhau, hoặc hai thành phần của dữ liệu ban đầu chênh lệch nhau quá lớn, lúc này ta cần chuẩn hóa dữ liệu trước khi thực hiện các bước tiếp theo.

2.2.1 Chuyển khoảng giá trị

Phương pháp đơn giản nhất là đưa tất cả các đặc trưng về cùng 1 khoảng, ví dụ [0, 1] hoặc [-1, 1]. Để đưa đặc trưng thứ i của một vector đặc trưng x về khoảng [0, 1], ta sử dụng công thức:

$$x'_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (2-1)$$

Trong đó x_i và x'_i lần lượt là giá trị đặc trưng ban đầu và giá trị đặc trưng sau khi được chuẩn hóa. $\min(x_i)$, $\max(x_i)$ là giá trị nhỏ nhất và lớn nhất của đặc trưng thứ I xét trên toàn bộ tập dữ liệu huấn luyện.

2.2.2 Chuẩn hóa theo phân phối chuẩn

Một phương pháp khác thường được sử dụng là đưa mỗi đặc trưng về dạng một phân phối chuẩn có kỳ vọng là 0 và phương sai là 1. Công thức chuẩn hóa là:

$$x'_i = \frac{x_i - \bar{x}_i}{\sigma_i} \quad (2-2)$$

2.2.3 Chuẩn hóa về cùng norm

Một lựa chọn khác cũng được sử dụng rộng rãi là biến vector dữ liệu thành vector có độ dài Euclid bằng 1. Việc này có thể được thực hiện bằng cách chia mỗi vector cho l_2 norm của nó:

$$x' = \frac{x}{\|x\|_2} \quad (2-3)$$

2.2.4 Xử lý dữ liệu rỗng

Không phải lúc nào các vector đặc trưng cũng có đầy đủ dữ liệu. Thực tế ta thường phải xây dựng mô hình trên những vector đặc trưng tồn tại nhiều vị trí mà dữ liệu bị mất, lỗi hoặc trống rỗng. Có hai cách cơ bản xử lý vấn đề trên:

- Xóa hàng chứa dữ liệu rỗng.
- Tính toán một giá trị thay thế trung tính mà không ảnh hưởng đến kết quả học của mô hình

2.3 Các mô hình học máy hồi quy

Học máy là một kỹ thuật dựa trên xác suất thống kê và dữ liệu lớn mà không cần lập trình rõ ràng, do đó ta có thể tiết kiệm thời gian lập trình và tránh sa vào các điều kiện lý tưởng thiếu thực tế, đồng thời nó cũng có thể tự cải thiện độ chính xác của các dự đoán đầu ra. Mục tiêu chính của mọi thuật toán học máy là tìm ra một bộ tham số tối ưu để điều chỉnh giả thuyết phù hợp với một bộ dữ liệu cụ thể.

2.3.1 Hồi quy tuyến tính (Linear Regression) [7]

Hồi quy tuyến tính là một thuật toán hồi quy mà đầu ra là một hàm số tuyến tính của đầu vào. Đây là thuật toán đơn giản nhất trong nhóm các thuật toán học có giám sát.

2.3.1.1 Hàm dự đoán đầu ra

Tổng quát, nếu mỗi điểm dữ liệu được mô tả bởi một vector đặc trưng d chiều, $x \in \mathbb{R}^d$, hàm dự đoán đầu ra được viết dưới dạng:

$$y = w_1 x_1 + w_2 x_2 + \dots + w_d x_d = x^T w \quad (2-4)$$

2.3.1.2 Sai số dự báo

Sau khi xây dựng được mô hình dự đoán đầu ra, ta cần tìm một phép đánh giá phù hợp với bài toán. Với bài toán hồi quy nói chung, ta mong muốn nhận được sai số e giữa đầu ra thực sự y và đầu ra dự đoán \hat{y} là nhỏ nhất:

$$\frac{1}{2}e^2 = \frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y - x^T w)^2 \quad (2-5)$$

Ở đây, ta lấy bình phương e vì hiệu $y - \hat{y}$ có thể là một số âm. Ta có thể tính sai số nhỏ nhất của mô hình bằng cách lấy trị tuyệt đối $|e| = |y - \hat{y}|$. Tuy nhiên, hàm trị tuyệt đối không khả vi tại gốc tọa độ, không thuận tiện cho việc tối ưu nên không được sử dụng rỗng rãi. Hệ số $\frac{1}{2}$ sẽ bị triệt tiêu khi ta lấy đạo hàm của e theo tham số mô hình w.

2.3.1.3 Hàm mất mát

Việc tính toán sai số với tất cả các cặp dữ liệu $(x_i, y_i), i=1, 2, \dots, N$, với N là số lượng các cặp dữ liệu trong tập huấn luyện, cho ta hàm mất mát sau:

$$L(w) = \frac{1}{2N} \sum_{i=1}^N (y_i - x_i^T w)^2 \quad (2-6)$$

2.3.1.1 Nghiệm của hồi quy tuyến tính

Nhận thấy rằng hàm mất mát $L(w)$ có gradient tại mọi w . Giá trị tối ưu của w có thể tìm được thông qua việc giải phương trình đạo hàm của $L(w)$ theo w bằng 0:

$$\frac{\nabla L(w)}{\nabla w} = 0 \Leftrightarrow \frac{1}{N} X (X^T w - y) = 0 \Leftrightarrow X X^T w = X y \quad (2-7)$$

Nếu ma trận vuông $X X^T$ khả nghịch, phương trình 3.6 có nghiệm duy nhất:

$$w = (X X^T)^{-1} X y \quad (2-8)$$

Nếu ma trận $X X^T$ không khả nghịch, ta có:

$$w = (X X^T)^\dagger X y \quad (2-9)$$

Trong đó, $(X X^T)^\dagger$ là giả khả nghịch.

2.3.1.2 Hệ số điều chỉnh

Hàm dự đoán đầu ra của hồi quy tuyến tính thường có thêm một hệ số điều chỉnh (bias) b:

$$f(x) = x^T w + b \quad (2-10)$$

Nếu $b = 0$, đường thẳng/mặt phẳng/siêu phẳng $y = x^T w + b$ luôn đi qua gốc tọa độ, việc thêm hệ số b khiến mô hình linh hoạt hơn. Hệ số điều chỉnh này cũng là một tham số mô hình.

Để thấy rằng, nếu coi mỗi điểm dữ liệu có thêm 1 đặc trưng $x_0 = 1$, ta sẽ có:

$$y = x^T w + b = w_1 x_1 + w_2 x_2 + \dots + w_d x_d + b x_0 = (\bar{x})^T (\bar{w}) \quad (2-11)$$

Trong đó, $\bar{x} = [x_0, x_1, \dots, x_N]^T$ và $\bar{w} = [b, w_1, w_2, \dots, w_N]$, ta có nghiệm của bài toán tối thiểu của hàm mất mát:

$$\bar{w} = (\bar{X} \bar{X}^T)^\dagger \bar{X} y \quad (2-12)$$

2.3.1.3 Đánh giá mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính có ý tưởng khá đơn giản, dễ hiện thực. Tuy nhiên nó vẫn có một số hạn chế nhất định.

Hạn chế đầu tiên của hồi quy tuyến tính là nó rất nhạy cảm với nhiễu. Nếu trong tập dữ liệu huấn luyện của mô hình xuất hiện một hoặc một vài cặp dữ liệu có giá trị sai lệch đáng kể so với các giá trị trung bình thì kết quả của mô hình sẽ khác đi rất nhiều. Để xử lý tình huống này, ta có thể tìm và loại bỏ nhiễu trong quá trình tìm nghiệm.

Hạn chế thứ hai của mô hình hồi quy tuyến tính là nó không biểu diễn được các mô hình phức tạp. Mặc dù phương pháp này có thể được áp dụng cho các quan hệ phi tuyến giữa đầu vào và đầu ra, mối quan hệ này vẫn đơn giản hơn rất nhiều so với thực tế.

2.3.2 Máy véc – to hỗ trợ (Support Vector Machine)

2.3.2.1 Xây dựng bài toán tối ưu cho máy véc – to hỗ trợ [7]

Giả sử ta có 2 lớp dữ liệu được gán nhãn 1 và -1 có bộ dữ liệu là các cặp (đặc trưng, nhãn): $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Mục tiêu của ta là tìm ra một siêu phẳng có dạng $w^T x + b = 0$ để biểu diễn mặt phân chia của 2 lớp. Với mỗi cặp dữ liệu bất kỳ, ta có khoảng cách từ x_n đến mặt phân chia là $\frac{y_n(w^T x_n + b)}{\|w\|_2}$. Trong đó, y_n và $(w^T x_n + b)$ là 2 величина

cùng dương hoặc cùng âm, nghĩa là ta luôn đảm bảo tử số là một величина không âm. Để đảm mặt phân chia $w^T x + b$ chia đều khoảng cách từ nó đến các phần tử gần nhất của 2 lớp, ta có khái niệm về lề (margin):

$$\text{margin} = \min_n \frac{y_n(w^T x_n + b)}{\|w\|_2} \quad (2-13)$$

Vì lề càng lớn thì khả năng phân loại sai dữ liệu càng thấp, do đó ta đưa bài toán tìm siêu phẳng $w^T x + b$ về bài toán tìm lề lớn nhất:

$$(w, b) = \arg \max_{w,b} \left\{ \min_n \frac{y_n(w^T x_n + b)}{\|w\|_2} \right\} = \arg \max_{w,b} \left\{ \frac{1}{\|w\|_2} \min_n y_n(w^T x_n + b) \right\} \quad (2-14)$$

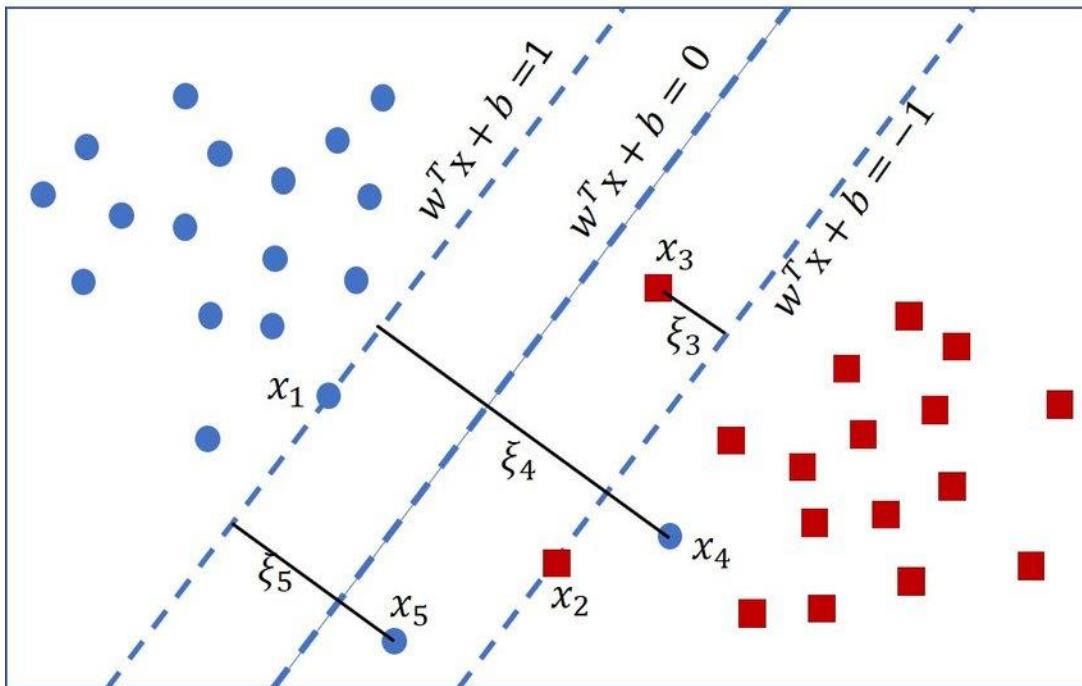
Vì khoảng cách từ các điểm dữ liệu đến mặt phân chia không đổi nên bài toán được đưa về dạng sau:

$$(w, b) = \arg \max_{w,b} \frac{1}{\|w\|_2} = \arg \min \frac{1}{2} \|w\|_2^2 \quad (2-15)$$

thỏa mãn: $1 - y_n(w^T x_n + b) \leq 0, \forall n = 1, 2, \dots, n$

Tuy nhiên, trong trường hợp 2 lớp dữ liệu tách biệt tuy nhiên nhưng tồn tại một vài điểm dữ liệu làm cho lề trở nên quá nhỏ và tăng tỉ lệ phân loại sai; hoặc thậm chí 2 lớp dữ liệu có một vài điểm dữ liệu đan xen lẫn nhau, trường hợp này vô nghiệm nếu ta tiếp cận

theo cách như trên. Để giải quyết bài toán này, ta cần hy sinh một vài điểm dữ liệu bằng cách chấp nhận cho chúng rơi vào vùng “không an toàn” chính là khoảng cách từ lề của lớp này đến lề của lớp kia. Việc đánh đổi này cũng cần được kiểm soát, tránh tình trạng hy sinh hầu hết các điểm, do đó bài toán của ta chuyển thành bài toán tối đa hóa lề và tối thiểu hóa sự hy sinh.



Hình 2-3: Mô hình máy véc – to hỗ trợ (lề mềm)

Với mỗi điểm dữ liệu x_n , ta thêm một biến slack (biến lỏng lẻo) ξ_n ; biến này bằng 0, tức là x_n nằm trong vùng an toàn, ngược lại nếu x_n nằm trong vùng không an toàn thì biến slack có giá trị > 0 để đo sự hy sinh. Đại lượng ξ_i tỉ lệ thuận với khoảng cách từ vị trí vi phạm đến lề của lớp tương ứng mà điểm dữ liệu này thuộc về. Và do đó, ξ_i được định nghĩa như sau:

$$\xi_i = |w^T x_i + b - y_i| \quad (2-16)$$

Do đó, phương trình (2-17) trở thành:

$$(w, b, \xi) = \arg \min_{w, b, \xi} \left(\frac{1}{2} \|w\|_2^2 + C \sum_{n=1}^N \xi_n \right) \quad (2-18)$$

Trong đó, C là một hằng số dương. Hằng số C được dùng để điều chỉnh tầm quan trọng giữa độ rộng lê và sự hy sinh. Phương trình (3-18) phải thỏa mãn: $w^T x_n + b \geq 1 - \xi_n$, hay $1 - \xi_n - w^T x_n + b \leq 0, \forall n = 1, 2, \dots, N$; và $-\xi_n \leq 0, \forall n = 1, 2, \dots, N$.

Để giải bài toán này, ta sẽ hướng đến việc giải bài toán đối ngẫu của nó. Trước tiên, ta cần kiểm tra tiêu chuẩn Slatter của bài toán tối ưu lồi, nếu thỏa mãn tiêu chuẩn Slatter nghĩa là đối ngẫu mạnh thỏa mãn và ta có thể tìm nghiệm tối ưu của bài toán tối ưu lồi thông qua hệ điều kiện KKT (Karush-Kuhn-Tucker).

Ta thấy rằng, với mọi $n = 1, 2, \dots, N$ và (w, b) , ta luôn có thể tìm được các số dương $\xi_n, n = 1, 2, \dots, N$, đủ lớn sao cho $y_n(w^T x_n + b) + \xi_n > 1, \forall n = 1, 2, \dots, N$. Vì vậy tồn tại điểm khả thi chặt cho bài toán và thỏa mãn tiêu chuẩn Slatter.

Ta có hàm Lagrange của bài toán (2-19) như sau:

$$L(w, b, \xi, \lambda, \mu) = \frac{1}{2} \|w\|_2^2 + C \sum_{n=1}^N \xi_n + \sum_{n=1}^N \lambda_n (1 - \xi_n - y_n(w^T x_n + b)) - \sum_{n=1}^N \mu_n \xi_n \quad (2-20)$$

Trong đó, $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_N]^T \geq 0$ và $\mu = [\mu_1, \mu_2, \dots, \mu_N]^T \geq 0$ là các biến đối ngẫu Lagrange. Hàm số đối ngẫu của bài toán tối ưu được biểu diễn như sau:

$$g(\lambda, \mu) = \min_{w, b, \xi} L(w, b, \xi, \lambda, \mu) \quad (2-21)$$

Với mỗi cặp (λ, μ) , ta quan tâm đến bộ ba (w, b, ξ) thỏa mãn điều kiện đạo hàm của hàm Lagrange bằng không:

$$\nabla_w L = 0 \Leftrightarrow w = \sum_{n=1}^N \lambda_n y_n x_n \quad (2-22)$$

$$\nabla_b L = 0 \Leftrightarrow \sum_{n=1}^N \lambda_n y_n = 0 \quad (2-23)$$

$$\nabla_{\xi_n} L = 0 \Leftrightarrow \lambda_n = C - \mu_n \quad (2-24)$$

Phương trình (2-25) cho thấy rằng ta chỉ cần quan tâm đến những cặp (λ, μ) sao cho $\lambda_n = C - \mu_n$. Từ đó, ta suy ra: $0 \leq \lambda_n, \mu_n \leq C, n = 1, 2, \dots, N$. Thay các biểu thức vào hàm Lagrange, ta thu được hàm mục tiêu của hàm đối ngẫu:

$$g(\lambda, \mu) = \sum_{n=1}^N \lambda_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m y_n y_m x_n^T x_m \quad (2-26)$$

Phương trình (2-27) có thể thu gọn thành:

$$\lambda = \arg \max_{\lambda} g(\lambda) \quad (2-28)$$

Thỏa mãn: $\sum_{n=1}^N \lambda_n y_n = 0, 0 \leq \lambda_n \leq C, \forall n = 1, 2, \dots, N$.

Ta có hệ điều kiện KKT của bài toán tối ưu SVM (lè mềm):

$$1 - \xi_n - y_n (\mathbf{w}^T \mathbf{x}_n + b) \leq 0 \quad (2-29)$$

$$-\xi_n \leq 0 \quad (2-30)$$

$$\lambda_n \geq 0 \quad (2-31)$$

$$\mu_n \geq 0 \quad (2-32)$$

$$\lambda_n (1 - \xi_n - y_n (\mathbf{w}^T \mathbf{x}_n + b)) = 0 \quad (2-33)$$

$$\mu_n \xi_n = 0 \quad (2-34)$$

$$\mathbf{w} = \sum_{n=1}^N \lambda_n y_n \mathbf{x}_n \quad (2-35)$$

$$\sum_{n=1}^N \lambda_n y_n = 0 \quad (2-36)$$

$$\lambda_n = C - \mu_n \quad (2-37)$$

Từ điều kiện (2-38) và (2-39) ta thấy rằng chỉ những n ứng với $\lambda_n \geq 0$ mới đóng góp vào việc tính nghiệm w của bài toán. Tập hợp $S = \{n : \lambda_n > 0\}$ được gọi là tập hỗ trợ (support set) và $\{x_n, n \in S\}$ được gọi là tập các véc – tơ hỗ trợ. Khi $\lambda_n \geq 0$, ta có:

$$y_n (\mathbf{w}^T \mathbf{x}_n + b) = 1 - \xi_n \quad (2-40)$$

Khi $0 < \lambda_n < C$, các điểm x_n nằm chính xác trên 2 đường thẳng hỗ trợ (lè). Giá trị b được tính theo công thức:

$$b = \frac{1}{N_M} \sum_{m \in M} (y_m - \mathbf{w}^T \mathbf{x}_m) \quad (2-41)$$

Với $M = \{m : 0 < \lambda_m < C\}$ và N_M là số phần tử của S.

Nghiệm của bài toán SVM (lè mềm) được cho bởi 2 công thức sau:

$$\mathbf{w} = \sum_{m \in S} \lambda_m y_m \mathbf{x}_m \quad (2-42)$$

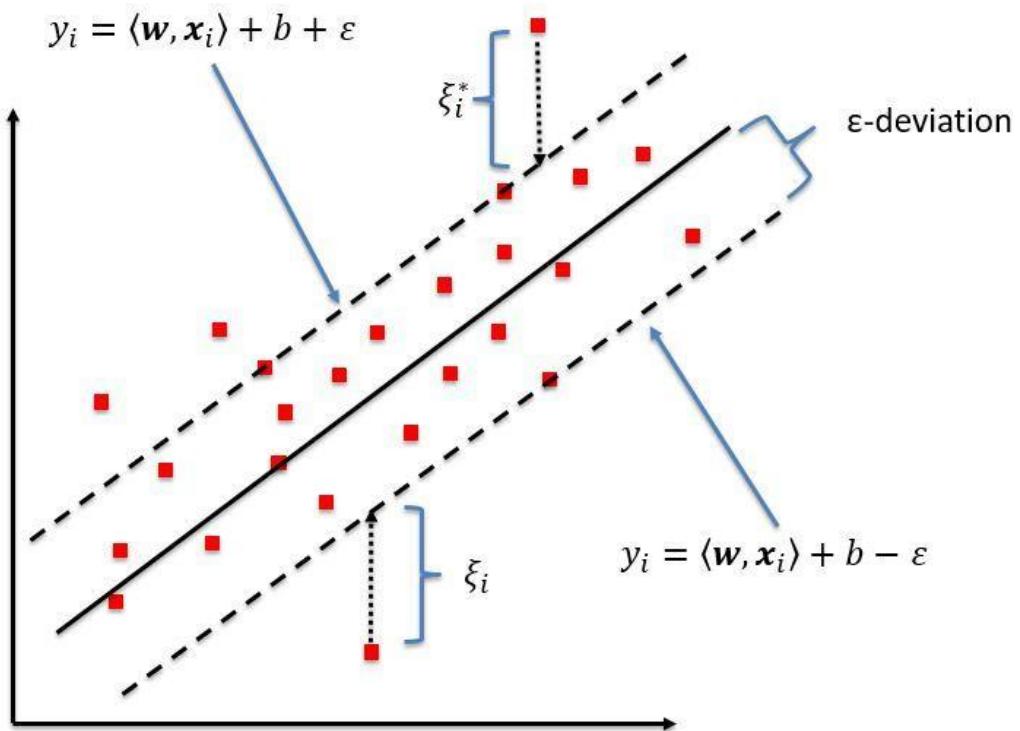
$$b = \frac{1}{N_M} \sum_{n \in M} (y_n - \mathbf{w}^T \mathbf{x}_n) = \frac{1}{N_M} \sum_{n \in M} \left(y_n - \sum_{m \in S} \lambda_m y_m \mathbf{x}_m^T \mathbf{x}_n \right) \quad (2-43)$$

2.3.2.2 Hồi quy véc – tơ hỗ trợ [8]

Cũng như bài toán phân loại, bài toán hồi quy của phương pháp máy véc – tơ hỗ trợ cũng đi tìm một đường thẳng/mặt phẳng/siêu phẳng có dạng $\mathbf{w}^T \mathbf{x} + b = 0$ đi qua các điểm dữ liệu sao cho với một độ rộng lè ε nhất định, tổng khoảng cách của mỗi điểm x_n đến siêu phẳng nhỏ hơn hoặc bằng ε . Tuy nhiên độ lớn của ε cũng cần được kiểm soát, vì khi dữ liệu phân tán mạnh, độ chính xác của mô hình trở nên không đáng tin cậy. Do vậy, ta cũng cần phải hy sinh một số điểm dữ liệu để đảm bảo ε vừa đủ lớn. Do đó, bài toán hồi quy véc – tơ hỗ trợ trở thành bài toán tối ưu:

$$(\mathbf{w}, b, \xi, \xi^*) = \arg \min_{\mathbf{w}, b, \xi, \xi^*} \left(\frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{n=1}^N (\xi_n + \xi_n^*) \right) \quad (2-44)$$

Trong đó, biến slack ξ là khoảng cách từ điểm dữ liệu x_n nằm ngoài vùng an toàn đến lề gần nhất là lề âm ($y_n = \mathbf{w}^T x_n + b - \xi$); ngược lại, ξ^* là khoảng cách ngắn nhất từ điểm dữ liệu x_n nằm ngoài vùng an toàn đến lề gần nhất là lề dương ($y_n = \mathbf{w}^T x_n + b + \xi^*$). Tương tự như trong bài toán phân loại, các giá trị biến slack ξ / ξ^* cũng phụ thuộc vào vị trí của điểm dữ liệu, những điểm nằm trong vùng an toàn thì $\xi = 0$; ngược lại, $\xi \geq 0$. Việc tối ưu hóa phương trình (2-45) cũng phải thỏa mãn các điều kiện: $y_n - \mathbf{w}^T x_n - b \leq \varepsilon + \xi_n$, $\mathbf{w}^T x_n + b - y_n \leq \varepsilon + \xi_n^*$ và $\xi_n, \xi_n^* \geq 0$.



Hình 2-4: Mô hình hồi quy vec – tơ hỗ trợ

2.3.2.3 Đánh giá mô hình máy véc – tơ hỗ trợ (Support Vector Machine)

Ưu điểm: Do bản chất của bài toán tối ưu lồi, lời giải sẽ luôn là tối thiểu toàn cục, không phải tối thiểu địa phương. SVM có thể được sử dụng cho dữ liệu tách biệt tuyến tính cũng như dữ liệu không tách biệt tuyến tính. SVM có thể được áp dụng cho các tập dữ liệu mà dữ liệu được gắn nhãn hoặc không được gắn nhãn (học bán giám sát). Ngoài ra, SVM có thể tính toán hiệu quả trên không gian đa chiều và tiết kiệm bộ nhớ do chỉ có một tập con của tập dữ liệu được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới, nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.

Nhược điểm: Trong trường hợp số lượng thuộc tính của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu thì SVM cho kết quả khá tệ. Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào.

2.4 Mạng nơ – ron hồi quy

Học sâu (còn được gọi là học có cấu trúc sâu) là một phần của họ các phương pháp học máy rộng hơn dựa trên mạng nơ – ron nhân tạo với học đại diện. Nó sử dụng nhiều lớp để trích xuất dần dần các đặc trưng cấp cao hơn từ đầu vào thô. Việc học có thể được giám sát, bán giám sát hoặc không giám sát.

Các kiến trúc học sâu như deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks và convolutional neural networks đã được áp dụng cho các lĩnh vực bao gồm thị giác máy tính, nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, dịch máy, tin sinh học, thiết kế thuốc, y tế phân tích hình ảnh, kiểm tra vật liệu và các chương trình trò chơi trên bàn cờ, trong đó chúng đã tạo ra kết quả có thể so sánh được và trong một số trường hợp vượt qua các chuyên gia con người.

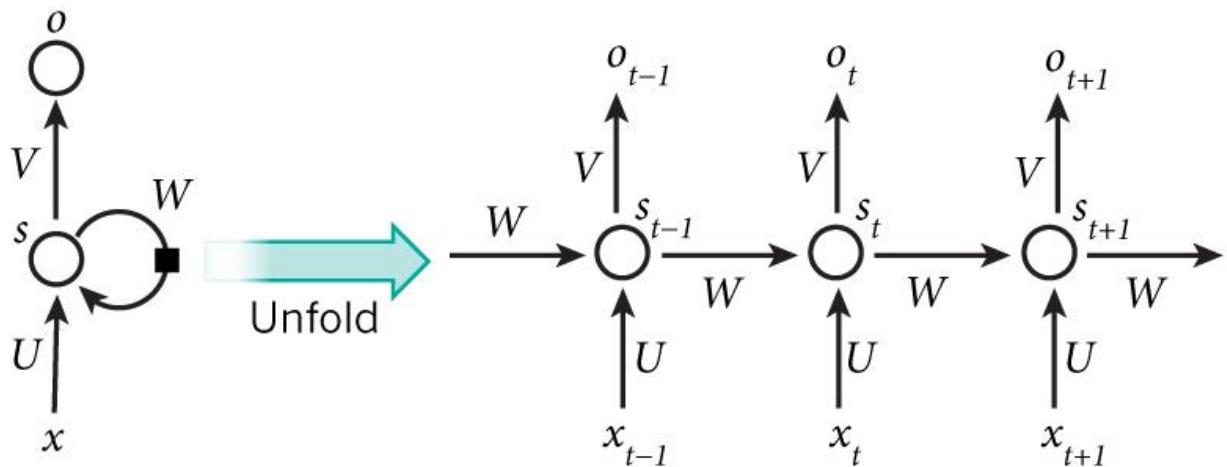
2.4.1 Mạng nơ – ron hồi quy (Recurrent Neural Network) [9]

Mạng nơ – ron hồi quy (RNN) là một lớp mạng nơ – ron nhân tạo trong đó các kết nối giữa các nút tạo thành một đồ thị có hướng theo trình tự thời gian. Điều này cho phép nó thể hiện hành vi động tạm thời. Bắt nguồn từ mạng nơ – ron truyền thống, RNN có thể sử dụng trạng thái bên trong (bộ nhớ) của chúng để xử lý chuỗi đầu vào có độ dài thay đổi.

Trong các mạng nơ – ron truyền thống, tất cả các đầu vào và cả đầu ra không liên kết thành chuỗi và độc lập với nhau. Nhưng các mô hình này không phù hợp trong rất nhiều bài toán. Ví dụ, nếu muốn dự đoán giá trị tiếp theo trong một chuỗi các giá trị có thứ tự thời gian thì ta cần phải biết các giá trị trước đó. RNN được gọi là mạng nơ – ron hồi quy bởi lẽ chúng thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó. Trên lý thuyết, RNN có thể sử dụng được thông tin của một chuỗi thông tin dài, tuy nhiên thực tế thì nó chỉ có thể nhớ được một vài bước trước đó.

2.4.1.1 Kiến trúc mạng nơ-ron hồi quy

Mạng nơ – ron hồi quy cho phép đầu ra của một lớp được sử dụng như đầu vào của lớp kế tiếp. Cụ thể như sau:



Hình 2-5: Kiến trúc mạng nơ – ron hồi quy

Tại mỗi bước t , giá trị kích hoạt h_t và đầu ra y_t (hay o_t) được biểu diễn như sau:

$$h_t = \sigma_h (W_h x_t + U_h h_{t-1} + b_h) \quad (2-46)$$

$$y_t = \sigma_y (W_y h_t + b_y) \quad (2-47)$$

Trong đó: x_t : vector đầu vào

h_t : vector lớp ẩn

y_t : vector đầu ra

W , U và b : vector và ma trận tham số mô hình

σ_h , σ_y : hàm kích hoạt

2.4.1.2 Hàm mất mát

Trong trường hợp của mạng neural hồi quy, hàm mất mát L của tất cả các bước thời gian được định nghĩa dựa theo mất mát ở mọi thời điểm như sau:

$$L(y, \hat{y}) = \sum_{t=1}^{T_y} L(y_t, \hat{y}_t) \quad (2-48)$$

2.4.1.3 Hàm lan truyền ngược theo thời gian

Lan truyền ngược được hoàn thành ở mỗi một thời điểm cụ thể. Ở bước T, đạo hàm của hàm mất mát L với ma trận trọng số W được biểu diễn như sau:

$$\frac{\partial L^T}{\partial W} = \sum_{t=1}^T \left. \frac{\partial L^T}{\partial W} \right|_{(t)} \quad (2-49)$$

2.4.1.4 Đánh giá kiến trúc mạng RNN

Ưu điểm:

- Khả năng xử lý đầu vào với bất kì độ dài nào
- Kích cỡ mô hình không tăng theo kích cỡ đầu vào
- Quá trình tính toán sử dụng các thông tin cũ
- Trọng số được chia sẻ trong suốt thời gian huấn luyện

Nhược điểm:

- Tính toán chậm
- Khó để truy cập các thông tin từ một khoảng thời gian dài trước đây
- Không thể xem xét bất kì đầu vào sau nào cho trạng thái hiện tại

2.4.2 Bộ nhớ ngắn hạn dài (Long Short – Term Memory) [9]

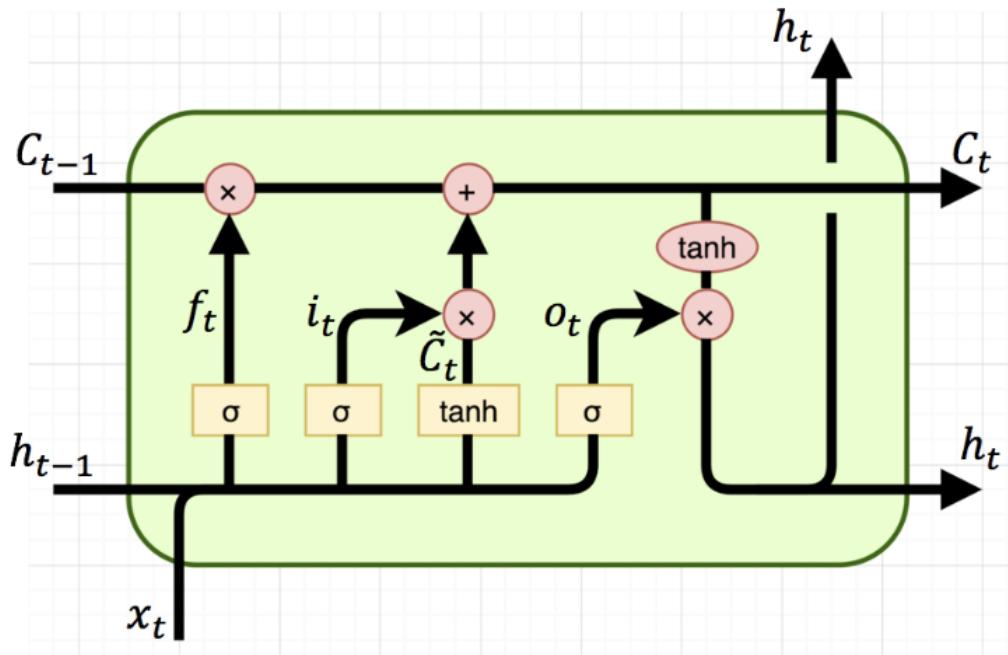
Bộ nhớ ngắn hạn dài (LSTM) là một kiến trúc mạng nơ – ron hồi quy (RNN) được sử dụng trong lĩnh vực học sâu. Không giống như các mạng nơ – ron truyền thống tiêu chuẩn, LSTM có các kết nối phản hồi. Nó không chỉ có thể xử lý các điểm dữ liệu đơn lẻ, mà còn toàn bộ chuỗi dữ liệu.

Một đơn vị LSTM thông thường bao gồm một ô nhớ, một cổng vào, một cổng ra và một cổng quên. Tế bào ghi nhớ các giá trị trong khoảng thời gian tùy ý và ba cổng điều chỉnh luồng thông tin vào và ra ô nhớ.

Về lý thuyết, các RNN cổ điển có thể theo dõi các phụ thuộc dài hạn tùy ý trong các trình tự đầu vào. Vấn đề với các RNN cổ điển có bản chất là khi huấn luyện một RNN cổ điển bằng cách truyền ngược (back – propagation), các gradient dài hạn được truyền ngược có thể biến mất (vanishing gradient problem), nghĩa là chúng có thể có xu hướng bằng không; hoặc bùng nổ (exploding gradient problem), nghĩa là chúng có thể có xu hướng tiến đến vô cùng, do các phép tính liên quan đến quá trình huấn luyện sử dụng các số có độ chính xác hữu hạn gây ra. LSTM có thể giải quyết một phần vấn đề vanishing gradient, tuy nhiên, nó vẫn có thể gặp phải vấn đề exploding gradient.

2.4.2.1 Cổng đầu vào, cổng quên và cổng đầu ra

Dữ liệu được đưa vào các cổng LSTM là đầu vào ở bước thời gian hiện tại X_t , và trạng thái ẩn ở bước thời gian trước đó H_{t-1} . Những đầu vào này được xử lý bởi một tầng kết nối đầy đủ và một hàm kích hoạt để tính toán các giá trị của các cổng đầu vào, cổng quên và cổng đầu ra. Kết quả là, tất cả các giá trị đầu ra tại ba cổng đều nằm trong khoảng $[0,1]$.

**Hình 2-6:** Một ô nhớ LSTM điển hình

Chúng ta giả sử rằng có h nút ẩn, mỗi minibatch có kích thước n và kích thước đầu vào là d . Như vậy, đầu vào là $X_t \in \mathbb{R}^{n \times d}$ và trạng thái ẩn của bước thời gian trước đó là $H_{t-1} \in \mathbb{R}^{n \times h}$. Tương tự, các công式 được định nghĩa như sau: công式 đầu vào là $i_t \in \mathbb{R}^{n \times h}$, công式 quên là $f_t \in \mathbb{R}^{n \times h}$, và công式 đầu ra là $o_t \in \mathbb{R}^{n \times h}$. Chúng được tính như sau:

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2-50)$$

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (2-51)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (2-52)$$

Trong đó $W_i, W_f, W_o \in \mathbb{R}^{d \times h}$ và $U_i, U_f, U_o \in \mathbb{R}^{h \times h}$ là các trọng số và $b_i, b_f, b_o \in \mathbb{R}^{1 \times h}$ là các hệ số điều chỉnh.

2.4.2.2 Vector kích hoạt đầu vào của ô nhớ

Bước tiếp theo là quyết định thông tin mới sẽ được lưu trữ ở trạng thái ô nhớ. Đầu tiên, một lớp sigmoid được gọi là lớp cổng đầu vào quyết định những giá trị nào được cập nhật. Tiếp theo, một lớp tanh tạo ra một vectơ có các giá trị mới, c_t , có thể được thêm vào trạng thái ô.

$$c_t = \sigma_c \left(W_c x_t + U_c h_{t-1} + b_c \right) \quad (2-53)$$

2.4.2.3 Vector trạng thái ô nhớ

Để cập nhật trạng thái mới, ta sẽ nhân trạng thái cũ với f_t để bỏ đi những thông tin ta quyết định quên lúc trước. Sau đó cộng thêm $i_t \times \tilde{C}_t$. Trạng thái mới thu được này phụ thuộc vào việc ta quyết định cập nhập mỗi giá trị trạng thái ra sao.

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{C}_t \quad (2-54)$$

2.4.2.4 Các trạng thái ẩn

Cuối cùng, chúng ta cần phải xác định cách tính trạng thái ẩn $h_t \in \mathbb{R}^{n \times h}$. Đây là nơi công đầu ra được sử dụng. Trong LSTM, đây chỉ đơn giản là một phiên bản có kiểm soát của hàm kích hoạt tanh trong ô nhớ. Điều này đảm bảo rằng các giá trị của h_t luôn nằm trong khoảng $[-1, 1]$. Bất cứ khi nào giá trị của cổng đầu ra là 1, thực chất chúng ta đang đưa toàn bộ thông tin trong ô nhớ tới bộ dự đoán. Ngược lại, khi giá trị của cổng đầu ra là 0, chúng ta giữ lại tất cả các thông tin trong ô nhớ và không xử lý gì thêm.

$$h_t = o_t \times \tanh(c_t) \quad (2-55)$$

2.4.3 Nút hồi quy có cổng (Gated Recurrent Unit) [10]

Các nút hồi quy có cổng (GRU) là một biến thể của mạng nơ – ron hồi quy, được giới thiệu vào năm 2014 bởi Kyunghyun Cho và cộng sự. Giống như LSTM, GRU có cổng quên, nhưng có ít tham số hơn LSTM, vì nó thiếu cổng ra. Hiệu suất của GRU trong một số nhiệm vụ về mô hình âm nhạc đa âm, mô hình tín hiệu giọng nói và xử lý ngôn ngữ tự nhiên được nhận thấy là tương tự như của LSTM. GRU đã được chứng minh là có hiệu suất tốt hơn trên một số tập dữ liệu có độ lớn và tần suất của dữ liệu nhỏ hơn.

2.4.3.1 Nút hồi quy có cổng đầy đủ

Có hai phiên bản của nút hồi quy có cổng, một là nút có cổng đầy đủ, một là nút có cổng rút gọn. Với nút có cổng đầy đủ ta có kiến trúc cơ bản sau:

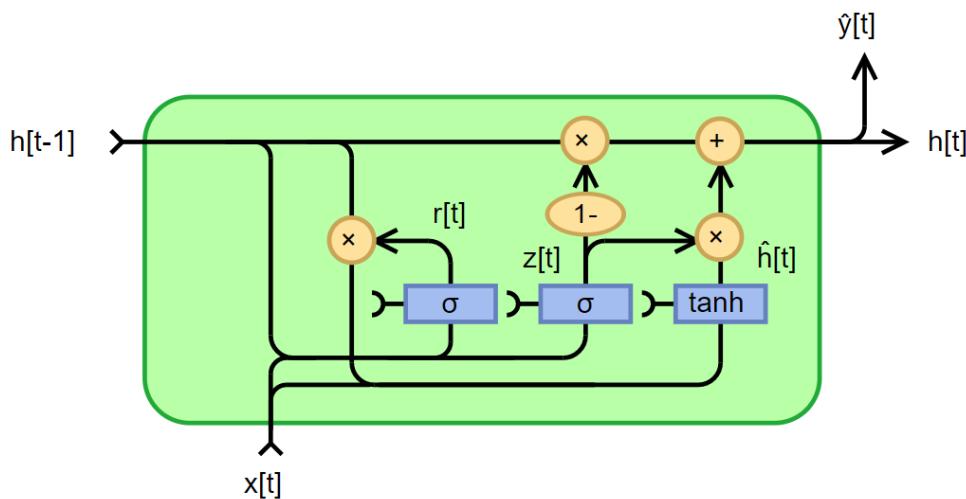
$$z_t = \sigma_g (W_z x_t + U_z h_{t-1} + b_z) \quad (2-56)$$

$$r_t = \sigma_g (W_r x_t + U_r h_{t-1} + b_r) \quad (2-57)$$

$$\hat{h}_t = \phi_h (W_h x_t + U_h (r_t \times h_{t-1}) + b_h) \quad (2-58)$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \hat{h}_t \quad (2-59)$$

Trong đó: x_t là vector đầu vào; h_t là vector đầu ra; \hat{h}_t là vector kích hoạt; z_t là vector cổng cập nhật; r_t là vector cổng cài đặt lại; W , U và b là những tham số mô hình; σ_g là hàm kích hoạt sigmoid; ϕ_h là hàm kích hoạt tanh.



Hình 2-7: Nút hồi quy có cổng đầy đủ điển hình

Ta có thể thay đổi kiến trúc của nút có cổng đầy đủ bằng cách thay đổi z_t và r_t . Dưới đây là ba biến thể khác của nút có cổng đầy đủ.

Loại 1: mỗi cổng chỉ phụ thuộc vào trạng thái ẩn trước đó và độ lệch.

$$z_t = \sigma_g(W_z h_{t-1} + b_z) \quad (2-60)$$

$$r_t = \sigma_g(U_r h_{t-1} + b_r) \quad (2-61)$$

Loại 2: Mỗi cổng chỉ phụ thuộc vào trạng thái ẩn trước đó.

$$z_t = \sigma_g(U_z h_{t-1}) \quad (2-62)$$

$$r_t = \sigma_g(U_r h_{t-1}) \quad (2-63)$$

Loại 3: Mỗi cổng được tính toán chỉ bằng cách sử dụng độ lệch.

$$z_t = \sigma_g(b_z) \quad (2-64)$$

$$r_t = \sigma_g(b_r) \quad (2-65)$$

2.4.3.2 Nút hồi quy có cỗng rút gọn

Nút hồi quy có cỗng rút gọn tương tự như phiên bản đầy đủ, ngoại trừ vector cỗng cập nhật và cài đặt lại được hợp nhất thành một cỗng quên. Điều này cũng có nghĩa rằng phương trình cho vectơ đầu ra phải được thay đổi:

$$f_t = \sigma_g \left(W_f x_t + U_f h_{t-1} + b_f \right) \quad (2-66)$$

$$\hat{h}_t = \phi_h \left(W_h x_t + U_h (r_t \times h_{t-1}) + b_h \right) \quad (2-67)$$

$$h_t = (I - f_t) \times h_{t-1} + f_t \times \hat{h}_t \quad (2-68)$$

Trong đó: x_t là vectơ đầu vào; h_t là vectơ đầu ra; \hat{h}_t là vectơ kích hoạt; f_t là vector cỗng quên; W , U và b là những tham số mô hình.

2.4.4 Các hàm kích hoạt thông dụng [11]

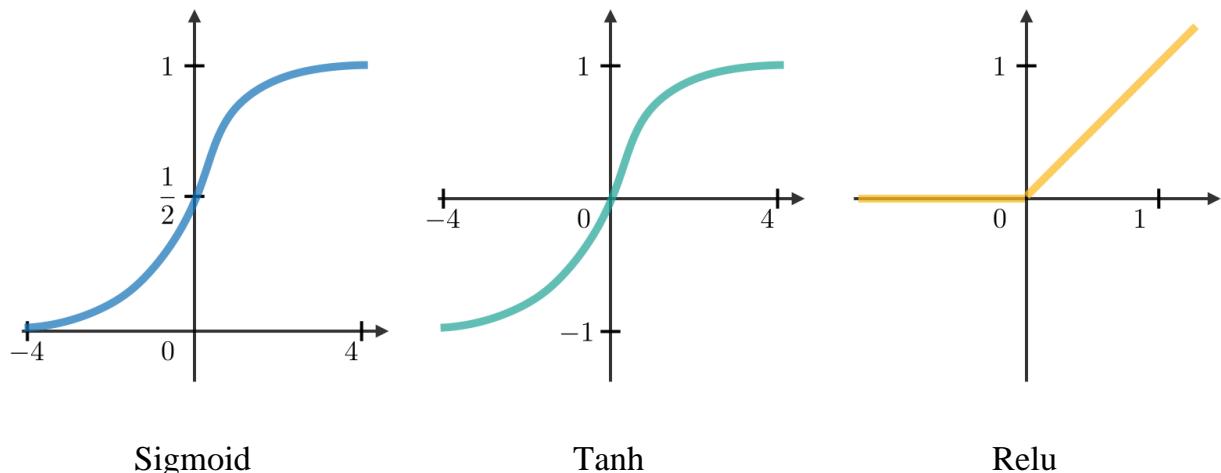
Hàm kích hoạt (activation function) mô phỏng tỷ lệ truyền xung qua axon của một nơ – ron thần kinh, là những hàm phi tuyến được áp dụng để chuẩn hóa đầu ra của các nơ – ron trong tầng ẩn của một mô hình mạng và được sử dụng làm đầu vào cho tầng tiếp theo.

Nếu không có các hàm kích hoạt, khả năng dự đoán của mạng sẽ bị giới hạn, sự kết hợp của các hàm kích hoạt giữa các tầng ẩn là để giúp mô hình học được các quan hệ phi tuyến phức tạp tiềm ẩn trong dữ liệu.

- Hàm Sigmoid: $Sigmoid(z) = \frac{1}{1 + e^{-z}}$ (2-69)

- Hàm Tanh: $Tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ (2-70)

- Hàm Relu: $Relu(z) = max(0, z)$ (2-71)



Hình 2-8: Các hàm kích hoạt thường dùng

2.4.5 Các hàm tối ưu hóa thông dụng [12]

Trong học máy nói riêng và các bài toán tối ưu nói chung, để tìm nghiệm cho bài toán ta phải tìm các điểm cực tiểu toàn cục của hàm số. Xét riêng các hàm khả vi, việc giải phương trình đạo hàm bằng không có thể vô cùng phức tạp hoặc có vô số nghiệm. Để đơn giản hóa, người ta tìm các điểm cực tiểu cục bộ và coi đó là nghiệm của bài toán trong trường hợp cụ thể.

Nếu tìm được một tập hữu hạn các điểm cực tiểu địa phương, ta chỉ cần lần lượt thay vào hàm số để tìm nghiệm tối ưu. Tuy nhiên trong hầu hết các trường hợp, việc giải phương trình đạo hàm bằng không là bất khả thi. Một hướng tiếp cận khác để giải bài toán tối ưu là chọn một điểm xuất phát, sau đó tiến dần đến đích sau mỗi lần lặp.

2.4.5.1 Gradient descent

Gradient descent là giải thuật đơn giản nhất trong việc tìm nghiệm tối ưu cho các thuật toán học máy. Giả sử ta cần tìm cực tiểu toàn cục cho hàm $f(\theta)$, trong đó θ là tập các tham số cần tối ưu. Gradient của hàm số đó tại một điểm θ bất kỳ được ký hiệu là $\nabla_{\theta} f(\theta)$. Thuật toán gradient descent bắt đầu bằng một điểm dự đoán θ_0 , sau đó sử dụng quy tắc cập nhật:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} f(\theta_t) \quad (2-72)$$

2.4.5.2 Momentum

Việc tìm nghiệm tối ưu thông qua thuật toán gradient descent được ví như như một viên bi lăn từ đỉnh đồi xuống thung lũng. Vấn đề được đặt ra là làm sao để viên bi vượt qua các điểm cực tiểu địa phương để tiến về thung lũng sâu nhất. Ứng dụng vật lý, ta hoàn toàn có thể giải quyết bài toán trên bằng cách cấp đà cho nó.

Trong gradient descent, ta cần tính lượng thay đổi tại thời điểm t để cập nhật vị trí mới cho nghiệm. Nếu ta coi đại lượng này là vận tốc v_t , vị trí mới cho hòn bi sẽ là $\theta_{t+1} = \theta_t - v_t$, với giả sử rằng mỗi vòng lặp là một đơn vị thời gian. Dấu trừ thể hiện cho

việc phải di chuyển ngược với gradient. Đại lượng v_t được thiết kế sao cho vừa mang thông tin của gradient hiện tại, vừa mang thông tin của đà:

$$v_t = \gamma v_{t-1} + \eta \nabla_{\theta} f(\theta_t) \quad (2-73)$$

Trong đó γ là một số dương nhỏ hơn 1, thường là 0.8 – 0.9. Khi công thức nghiệm trở thành:

$$\theta_{t+1} = \theta_t - v_t = \theta_t - \eta \nabla_{\theta} f(\theta_t) - \gamma v_{t-1} \quad (2-74)$$

2.4.5.3 Root mean square propagation

RMSprop tìm tối ưu bằng cách sử dụng đường trung bình động của các gradient bình phương để chuẩn hóa gradient. Quá trình chuẩn hóa này cân bằng kích thước bước (độ lượng), giảm bước cho các gradient lớn để tránh phát nổ và tăng bước cho các gradient nhỏ để tránh biến mất. Quy tắc cập nhật như sau:

$$v_t = \gamma v_{t-1} + (1-\gamma) (\nabla_{\theta} f(\theta_{t-1}))^2 \quad (2-75)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t}} \nabla_{\theta} f(\theta_t) \quad (2-76)$$

2.4.5.4 Adaptive Moment Estimation

Tối ưu hóa kiểu Adaptive Moment Estimation (Adam) là một thuật toán kết hợp kỹ thuật của RMSprop và Momentum. Thuật toán sử dụng internal states momentum (m) và squared momentum (v) của gradient cho các tham số. Sau mỗi batch huấn luyện, giá trị của m và v được cập nhật như sau:

$$m_{\theta}^{(t+1)} = \beta_1 m_{\theta}^{(t)} + (1-\beta_1) \nabla_{\theta} L^{(t)} \quad (2-77)$$

$$v_{\theta}^{(t+1)} = \beta_2 m_{\theta}^{(t)} + (1-\beta_2) (\nabla_{\theta} L^{(t)})^2 \quad (2-78)$$

$$\hat{m}_\theta = \frac{m_\theta^{(t+1)}}{1 - \beta_1} \quad (2-79)$$

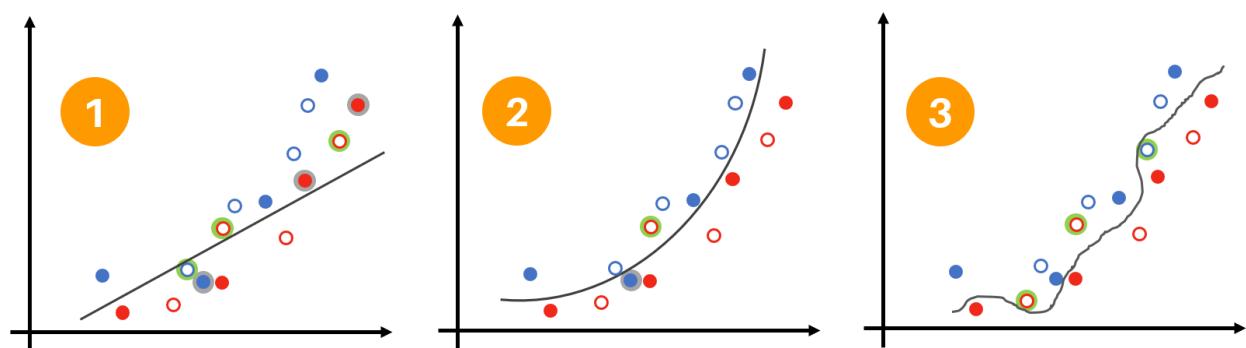
$$\hat{v}_\theta = \frac{v_\theta^{(t+1)}}{1 - \beta_2} \quad (2-80)$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\hat{m}_\theta}{\sqrt{\hat{v}_\theta}} + \varepsilon \quad (2-81)$$

Trong đó, L là hàm măt măt, ε là một giá trị nhỏ để tránh tình huống chia cho 0, β_1 và β_2 cũng giống như γ bên trên là những giá trị làm tăng tốc độ hội tụ.

2.5 Quá khớp

Trong các mô hình học có giám sát, ta thường phải đi tìm một mô hình ánh xạ các vector đặc trưng thành các kết quả tương ứng trong tập huấn luyện. Một cách tự nhiên, ta sẽ đi tìm các tham số mô hình sao cho việc xấp xỉ có sai số càng nhỏ càng tốt. Điều này nghĩa là mô hình càng khớp với dữ liệu càng tốt. Tuy nhiên, sự thật là nếu một mô hình quá khớp với dữ liệu huấn luyện thì nhiều khả năng nó sẽ không mang lại kết quả tốt trên tập kiểm tra. Nói cách khác, mô hình này không có tính tổng quát.



Hình 2-9: Các hiện tượng underfitting, fitting và overfitting

Như thể hiện trong hình 3-8, với biểu đồ 1 ta thấy rằng đường hồi quy không thực sự phản ánh đúng biểu hiện thực tế của tập dữ liệu, nghĩa là mô hình này chưa phải là một mô hình tối ưu, do đó ta gọi hiện tượng này là underfitting. Đối với đồ thị 2, đường cong gần như phản ánh đúng xu hướng của tập dữ liệu, đây là trường hợp mà ta mong muốn trong mọi bài toán học máy. Còn lại biểu đồ 3, đường cong uốn lượn quá nhiều, điều này có thể mang lại những kết quả khá quan trọng trên tập huấn luyện vì nó giữ sai số ở mức thấp, tuy nhiên việc này cũng làm mất đi tính tổng quát của bài toán, khi ta kiểm tra khả năng cao là mô hình sẽ đưa ra dự báo có độ sai lệch lớn với thực tế. Ta gọi trường hợp này là overfitting.

Trong cả 3 trường hợp, ta đặc biệt quan tâm overfitting. Việc tránh overfitting khá phức tạp, có nhiều chiến thuật tránh overfitting có thể kể đến như chia tập dữ liệu thành 3 phần để xác nhận và kiểm tra sau khi huấn luyện, chia tập dữ liệu thành nhiều tập con để xác thực chéo và kết thúc sớm quá trình huấn luyện để đảm bảo mô hình học vừa đủ.

Cách đơn giản nhất để tránh hiện tượng quá khớp là trích từ tập huấn luyện ra một tập con nhỏ và tiến hành đánh giá mô hình trên tập dữ liệu này. Tập dữ liệu này gọi là tập xác thực. Lúc này tập huấn luyện mới là phần còn lại của tập huấn luyện ban đầu.

Cũng với ý tưởng trích xuất một tập con nhỏ ra để đánh giá mô hình, nhưng trong nhiều trường hợp dữ liệu để huấn luyện mô hình là hạn chế, nếu trích quá nhiều dữ liệu để xác thực thì phần còn lại không đủ để xây dựng mô hình, còn nếu tập xác thực quá nhỏ thì có khả năng gây ra hiện tượng thiên lệch, không đánh giá đúng mô hình. Để xử lý tình huống trên, ta chia tập huấn luyện thành k tập con nhỏ, tại mỗi lần thử một tập con được lấy ra làm tập xác thực, k – 1 tập còn lại dùng cho việc huấn luyện. Như vậy ta có k mô hình với k sai số huấn luyện và k sai số xác thực, các sai số trung bình được tính là trung bình cộng của các sai số tương ứng. Cách làm này gọi là xác thực chéo.

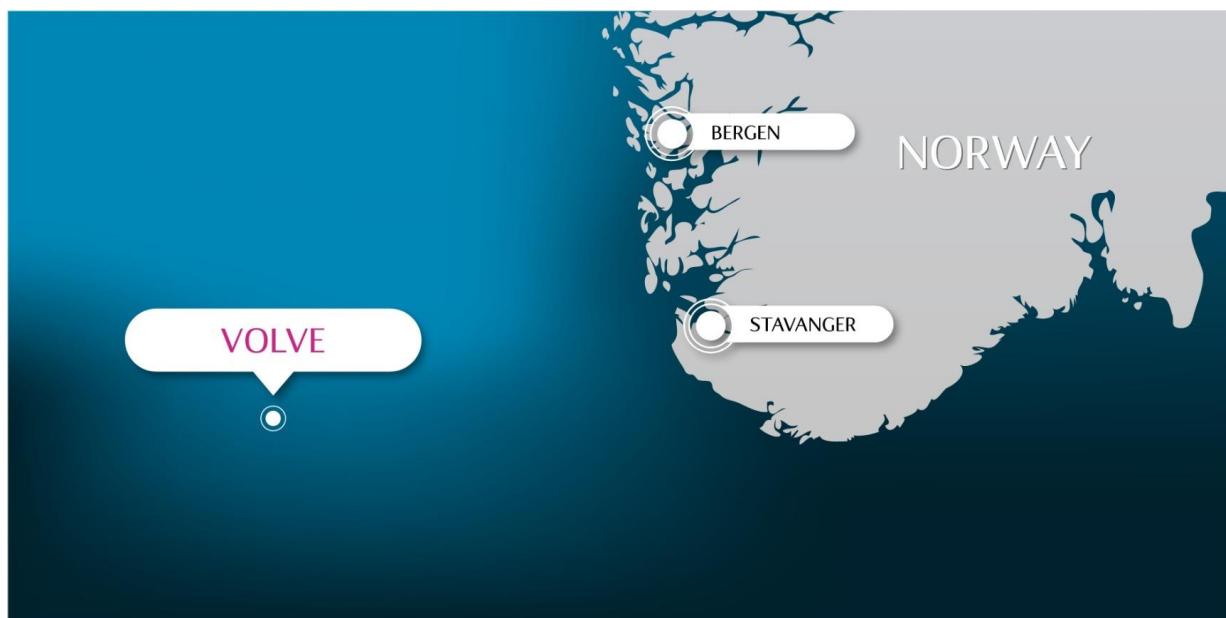
Một cách khác để giải quyết vấn đề quá khớp là khi huấn luyện ta có thể kết thúc sớm quá trình này. Khi huấn luyện, ta tính toán cả sai số huấn luyện và sai số xác thực, nếu sai số huấn luyện có xu hướng giảm nhưng sai số xác thực lại tăng thì ta kết thúc thuật toán.

CHƯƠNG 3: DỰ BÁO KHAI THÁC GIÉNG F14 – MỎ VOLVE

3.1 Tổng quan về mỏ Volve

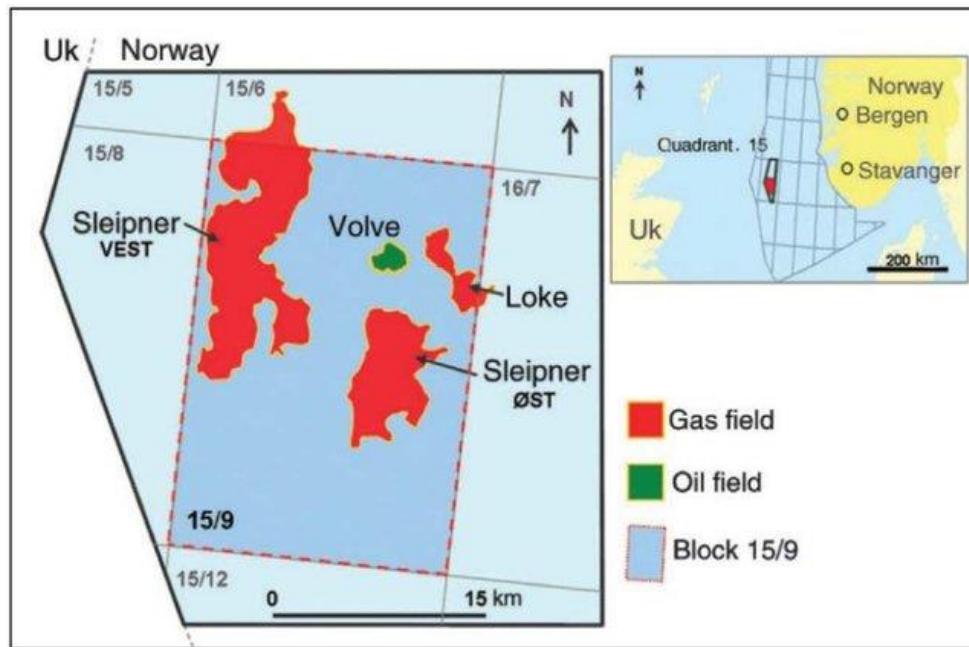
3.1.1 Mô tả mỏ

Volve là một mỏ đã ngừng hoạt động nằm ở Biển Bắc, được phát hiện vào năm 1993. Mỏ này nằm cách Stavanger 200 km về phía tây như trong hình 2.1 và cách mỏ Sleipner Ost 5 km về phía bắc với độ sâu 80m trong block 15/9. Việc khoan bắt đầu từ tháng 5 năm 2007, đi vào khai thác vào năm sau đó và kết thúc vào năm 2016 sau 8,5 năm hoạt động, gấp hơn hai lần so với kế hoạch ban đầu. Các giếng mới đang được khoan cho đến năm 2012 – 2013, góp phần làm tăng tỷ lệ thu hồi và kéo dài tuổi thọ của mỏ. Tuy nhiên, nguồn tài nguyên còn lại rất hạn chế và do giá dầu giảm trong những năm gần đây, các giếng mới không còn khả năng sinh lời. Tất cả các khả năng kéo dài tuổi thọ của mỏ đã được khám phá sau đó đều cho kết quả rất tốt.

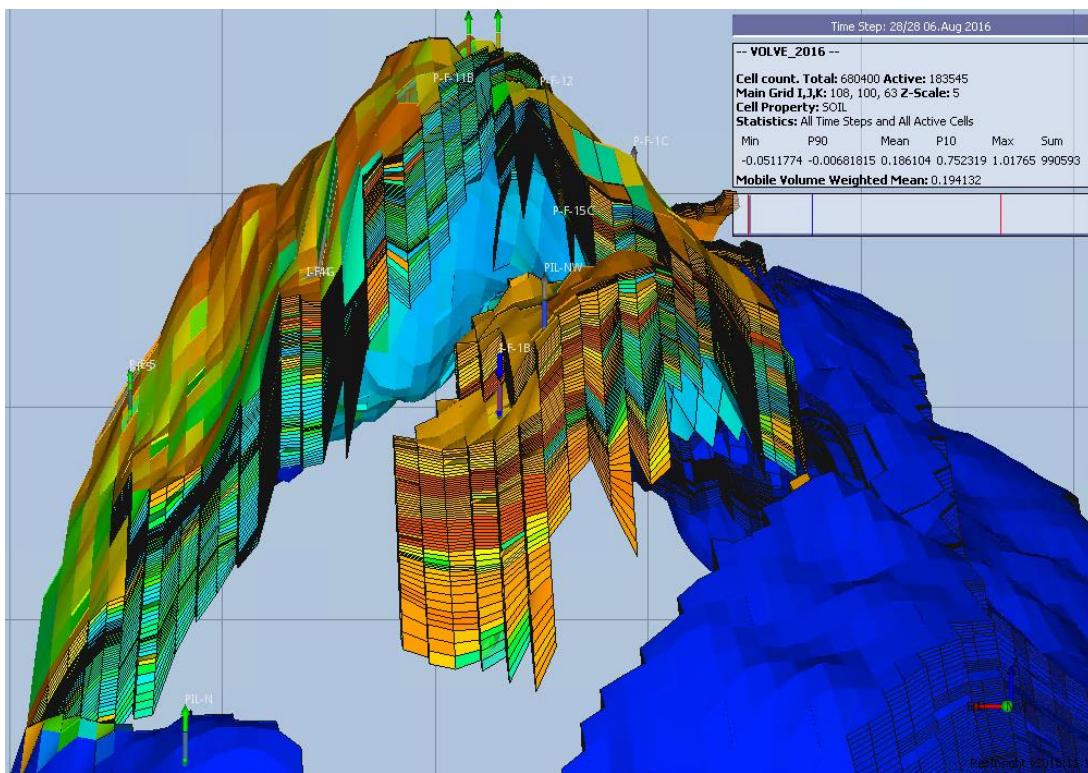


Hình 3-1: Vị trí của mỏ Volve

Volve khai thác dầu từ đá cát kết của kỷ Jura giữa trong hệ tầng Hugin. Vỉa ở độ sâu 2.700 – 3.100 mét. Phần phía tây của cấu trúc bị đứt gãy nặng và thông tin liên kết qua các đứt gãy là không chắc chắn. Mỏ được khai thác bằng phương pháp bơm ép nước để trợ áp.



Hình 3-2: Vị trí mỏ Volve trong block 15/9 và các mỏ lân cận

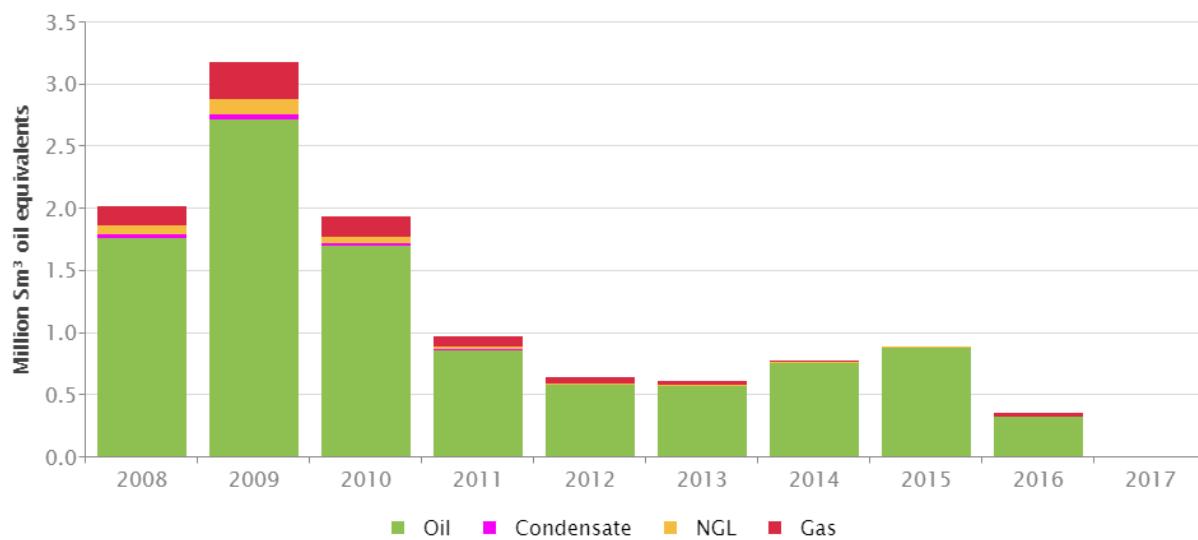


Hình 3-3: Mô hình vỉa của mỏ Volve

3.1.2 Lịch sử khai thác

Quá trình phát triển mỏ dựa trên hoạt động khai thác từ giàn jack – up Maersk Inspirer, với tàu Navion Sga được sử dụng làm tàu chứa dầu thô trước khi xuất khẩu. Khí được dẫn đến platform Sleipner để chế biến và xuất khẩu. Mỏ Volve đạt tỷ lệ thu hồi 54% và vào tháng 3 năm 2016, mỏ đã ngừng hoạt động vĩnh viễn. Ban đầu, lĩnh vực này được lên kế hoạch cho 3 – 5 năm hoạt động.

Bắt đầu từ tháng 2 năm 2008, quá trình khai thác mỏ Volve kéo dài khoảng 8 năm. Lúc cao điểm, mỏ khai thác 56.000 thùng mỗi ngày và tổng cộng 63 triệu thùng dầu được khai thác trước khi mỏ đóng cửa vào năm 2016. Mỏ này được phát triển khi giá dầu thấp và một mô hình phi truyền thống đã được áp dụng để thu hồi các nguồn tài nguyên một cách dễ dàng và có lợi nhuận. Dữ liệu hiện trường sẽ có mục đích mới là phục vụ cho học tập và nghiên cứu sau thời gian ngừng hoạt động. Những cá nhân và tổ chức được cấp phép cho Volve là ExxonMobil và Bayergas. Thông tin về địa chất, địa vật lý, khoan, mô hình tĩnh, mô phỏng động và các thông tin khác đã được cung cấp bởi công ty Equinor trong năm 2018. Sản lượng chi tiết từng năm được thể hiện trong hình bên dưới:



Hình 3-4: Dữ liệu khai thác trong giai đoạn 2008 – 2017

3.2 Quy trình thực hiện

Hầu hết mọi thuật toán học máy và học sâu đều tuân theo quy trình công việc bên dưới. Đầu tiên, ta thu thập tất cả dữ liệu thô mà bài toán yêu cầu. Tiếp theo đó ta tiến hành tiền xử lý dữ liệu, làm cho dữ liệu “sạch” hơn, phù hợp với đầu vào của các mô hình học máy, học sâu. Sau khi đã làm “sạch” dữ liệu, ta tiến hành lựa chọn và xử lý các đặc trưng sẽ được đưa vào mô hình.

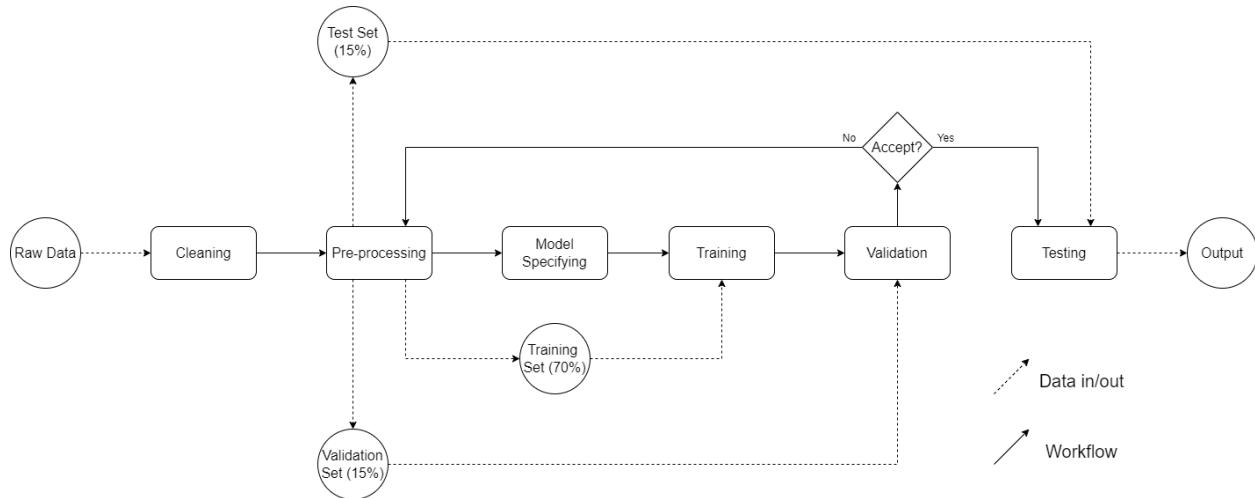
Dữ liệu được chia thành ba phần: dữ liệu huấn luyện (70%), dữ liệu xác thực (15%) và dữ liệu kiểm tra (15%). Trong kỹ thuật xây dựng đặc trưng đề cập bên trên, ta có 3 phương pháp hiệu chỉnh giá trị dữ liệu. Ta áp dụng phương pháp chuyển khoảng giá trị (min – max scaling) vì phân bố của dữ liệu là không đổi và nó cho kết quả tốt hơn các phương pháp khác.

Để xây dựng một mô hình học máy, học sâu hoàn thiện, trước tiên ta bắt đầu với những mô hình đơn giản nhất, gọi là bước khởi tạo mô hình. Khởi tạo mô hình có nghĩa là thiết lập tất cả các tham số của mô hình thành các giá trị cụ thể. Có hai cách phổ biến là khởi tạo tất cả các tham số bằng không hoặc thành các giá trị ngẫu nhiên từ phân phối chuẩn chuẩn hoặc phân phối đồng nhất và nhân nó với một đại lượng vô hướng.

Những mô hình đơn giản không phải lúc nào cũng cho ra kết quả tốt, nhất là đối với những bộ dữ liệu phức tạp. Chính vì thế, ta cần tối ưu hóa mô hình để tìm ra lời giải tốt hơn cho bài toán. Tối ưu hóa mô hình là tìm ra sự kết hợp tốt nhất giữa các siêu tham số của các mô hình nhằm mục đích thu được kết quả tốt nhất. Việc này đòi hỏi ta phải “thử và sai” nhiều lần mới có thể tìm ra lời giải tốt nhất.

Mô hình xác thực và kiểm tra được gọi là quá trình xử lý trong đó một mô hình được huấn luyện được đánh giá với một tập dữ liệu xác thực và kiểm tra. Chúng nhằm mục đích tìm ra mô hình tối ưu với hiệu quả cao nhất. Một số tiêu chí có thể được sử dụng để xác thực kết quả như sai số toàn phương trung bình, sai số tuyệt đối lớn nhất hoặc trung bình, hệ số xác định.

Sau khi hoàn thành tất cả các bước trước đó ở trên, mô hình có thể có khả năng dự báo. Trong nghiên cứu này, lưu lượng khai thác bao gồm lưu lượng dầu, lưu lượng khí là kết quả đầu ra từ mô hình đào tạo.



Hình 3-5: Quy trình chi tiết cho các thuật toán học máy và học sâu

3.3 Dữ liệu đầu vào

Bộ dữ liệu khai thác được sử dụng trong bài báo này là từ mỏ Volve trên thềm lục địa Na Uy với lịch sử khoảng 8 năm khai thác. Dữ liệu khai thác dầu, khí và nước hàng ngày đều được ghi lại bao gồm áp suất, nhiệt độ và độ mở ống góp. Các dữ liệu khai thác này là các biến đầu vào chính cho mô hình học máy và học sâu và nó đưa ra các dự đoán về khai thác nhiều giai đoạn trong tương lai dưới dạng đầu ra của mô hình. Trong nghiên cứu này, áp suất đáy (BTHP), áp suất dầu giếng (WHP), nhiệt độ dầu giếng (WHT), chênh áp suất trong ống chống (DP), độ mở ống góp (CS) theo tỷ lệ phần trăm là đầu vào để đào tạo thuật toán học máy và học sâu.

Quá trình thu thập dữ liệu phụ thuộc vào loại vấn đề cần giải quyết. Dữ liệu khai thác thường được tập hợp trong tệp excel gồm 24 cột dữ liệu và mỗi cột thể hiện một loại dữ liệu cụ thể. Do sự hỗn độn của dữ liệu bao gồm dữ liệu bị thiếu và bị nhiễu, trước khi áp dụng bất kỳ phương pháp học máy và học sâu nào, cần áp dụng kỹ thuật xây dựng đặc trưng để có được dữ liệu sạch và hữu dụng.

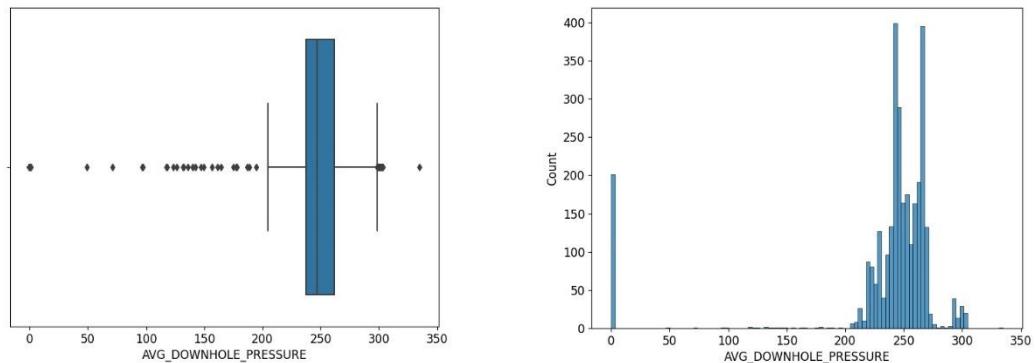
3.4 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một quá trình làm sạch dữ liệu thô. Một số loại dữ liệu bị thiếu, dữ liệu nhiễu và dữ liệu không nhất quán cần phải thông qua một bộ lọc thường được gọi là Feature Extraction hoặc kỹ thuật feature. Mặc dù có 24 cột dữ liệu thô nhưng hầu hết trong số đó là dữ liệu văn bản cung cấp thông tin về kiểu, ID, mã giếng và dữ liệu tên lưu và các dữ liệu không đầy đủ khác để áp dụng thuật toán ML/DL. Một khía cạnh quan trọng khác là giá trị nhiễu của dữ liệu. Có một số dữ liệu lỗi có thể có trong tập dữ liệu sai lệch đáng kể so với dữ liệu được quan sát. Vì vậy, loại dữ liệu này cũng cần được loại bỏ. Để xử lý vấn đề này, ta chỉ cần trích xuất những thông tin thống kê từ bộ dữ liệu ban đầu, sau đó ta tiến hành loại bỏ các hàng thiếu thông tin hoặc gây nhiễu cho tổng thể mô hình. Dữ liệu thô trước khi được xử lý được thống kê như sau:

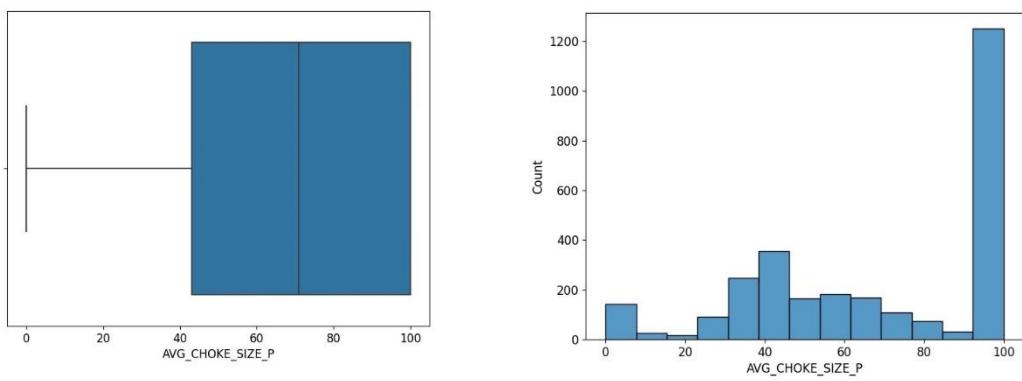
Bảng 3-1: Thống kê đơn giản của bộ dữ liệu thô trước khi xử lý

	BTHP	BTHT	DP	CS	WHP	WHT	OIL	GAS	WAT
count	3050,00	3050,00	3050,00	2860,00	3056,00	3056,00	3056,00	3056,00	3056,00
mean	233,07	95,13	192,65	69,39	41,53	77,10	1290,00	189139,25	2330,25
std	64,92	25,85	57,74	31,06	22,72	25,62	1298,36	184204,11	1462,93
min	0	0	0	0	0	0	0	0	-59,19
25%	237,49	99,62	180,75	43,01	31,02	81,22	209,86	31304,04	695,59
50%	246,78	101,01	204,25	71,01	33,55	86,96	880,79	142362,87	2965,72
75%	261,97	105,05	229,63	100,00	49,08	88,54	2033,70	305327,86	3444,15
max	334,66	106,77	302,11	100,00	125,72	93,51	5644,37	789974,73	5691,77

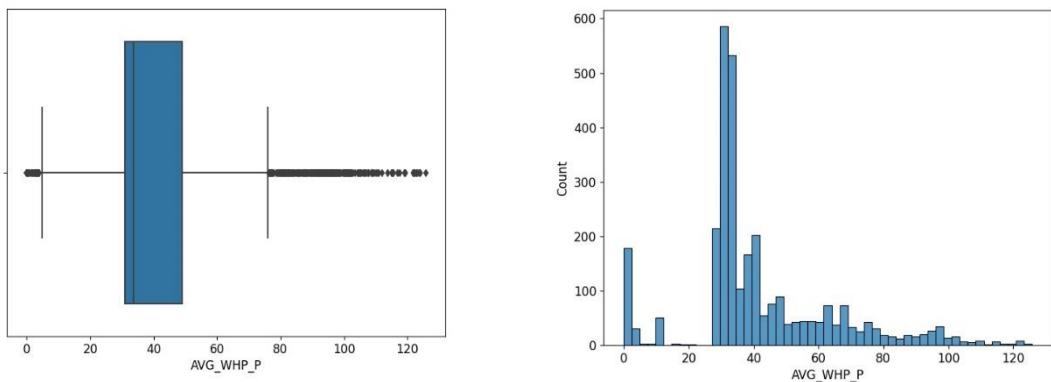
Các đặc trưng mà ta cần đặc biệt quan tâm là áp suất trung bình đáy giếng (BTHP), độ mở ống góp trung bình (CS), áp suất dầu giếng trung bình (WHP), lưu lượng trung bình của dầu, khí và nước (OIL, GAS, WAT), vì chúng có nhiều tác động trực tiếp và gián tiếp đến mô hình chung của ta. Dưới đây là các biểu đồ boxplot và histogram:



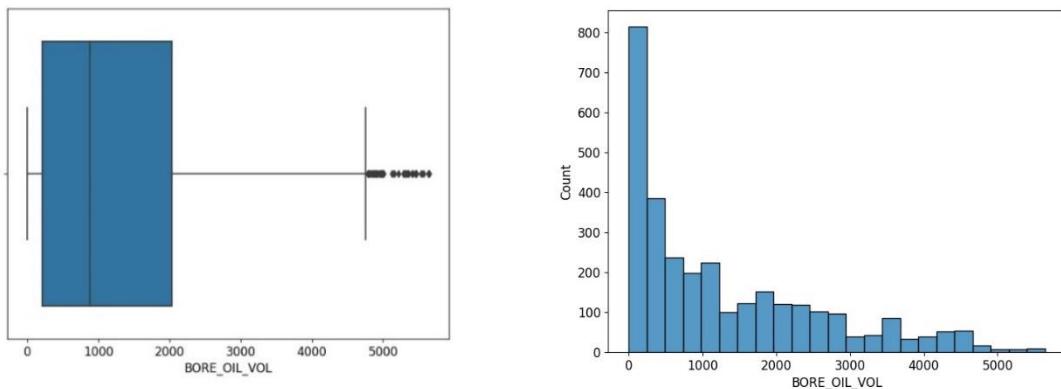
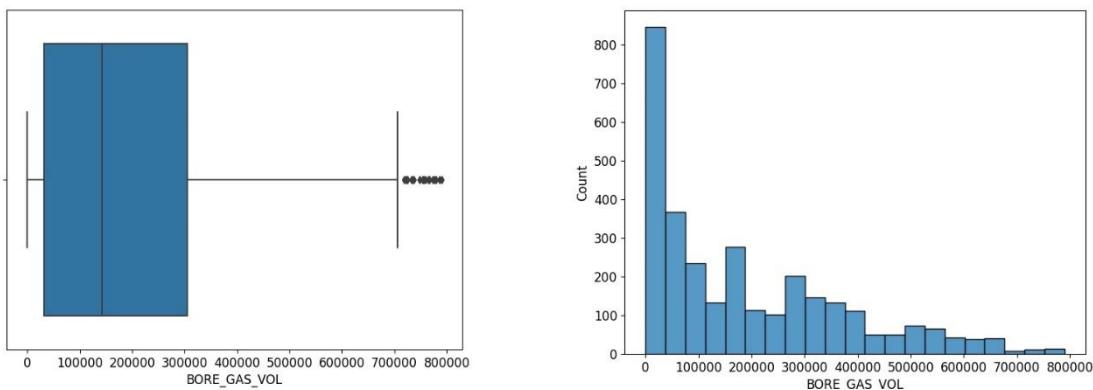
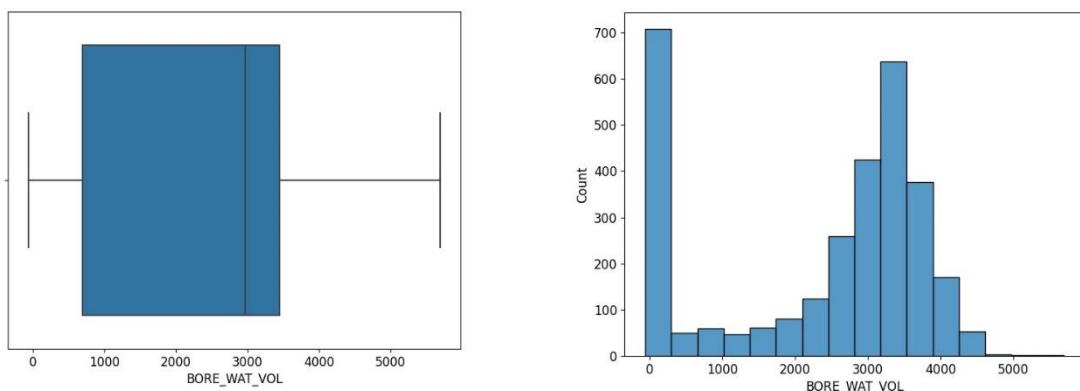
Biểu đồ 3-1: Biểu đồ boxplot và histogram của BTHP



Biểu đồ 3-2: Biểu đồ boxplot và histogram của CS



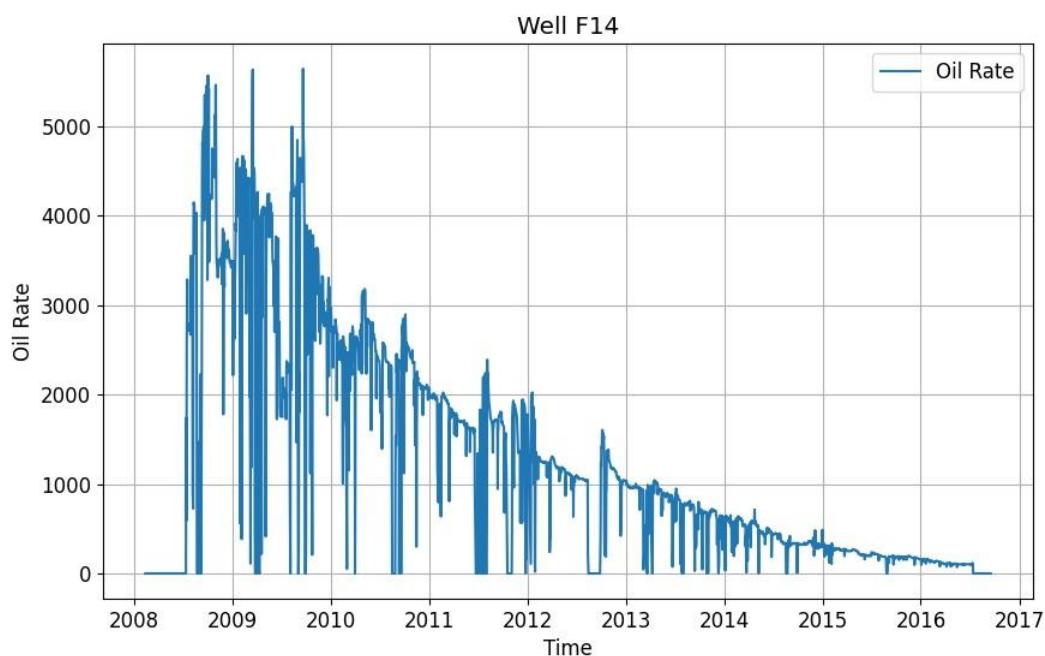
Biểu đồ 3-3: Biểu đồ boxplot và histogram của WHP

*Biểu đồ 3-4: Biểu đồ boxplot và histogram của OIL**Biểu đồ 3-5: Biểu đồ boxplot và histogram của GAS**Biểu đồ 3-6: Biểu đồ boxplot và histogram của WAT*

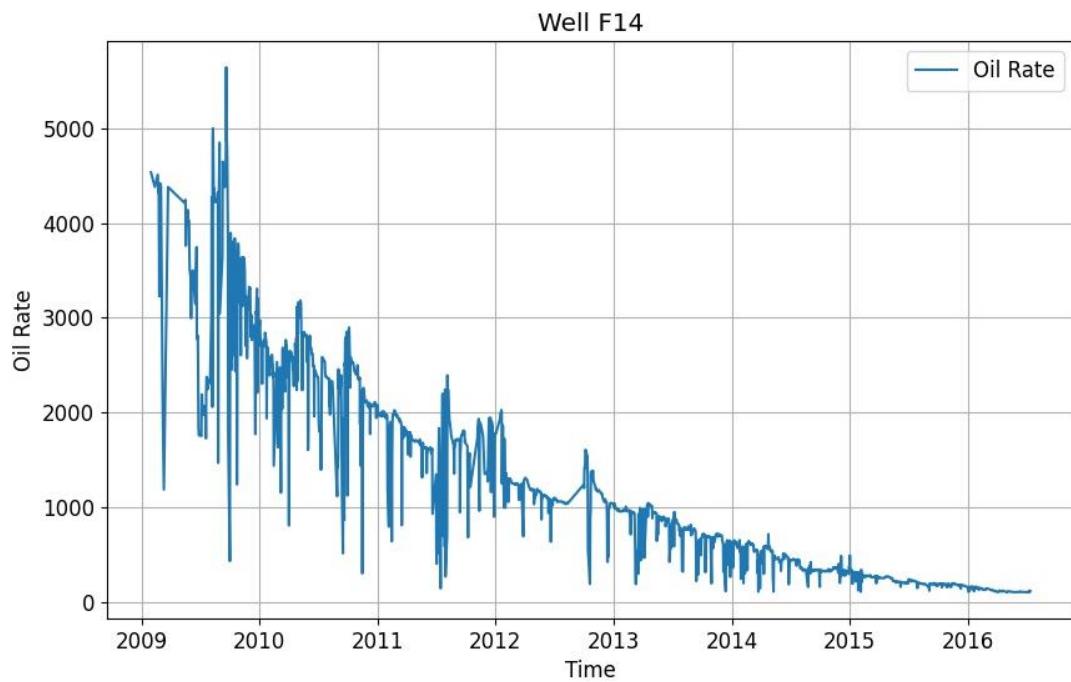
Sau khi phân tích, ta tiến hành loại bỏ các giá trị nhiễu. Ta loại bỏ các giá trị nhỏ hơn 20 đối với CS; tương tự, ta loại bỏ các giá trị nhỏ hơn 10 đối với WHP, các giá trị nhỏ hơn 100 đối với BTHP, OIL, GAS và WAT. Tiến hành thống kê lại, ta được bảng sau:

Bảng 3-2: Thống kê đơn giản về dữ liệu thô sau khi xử lý

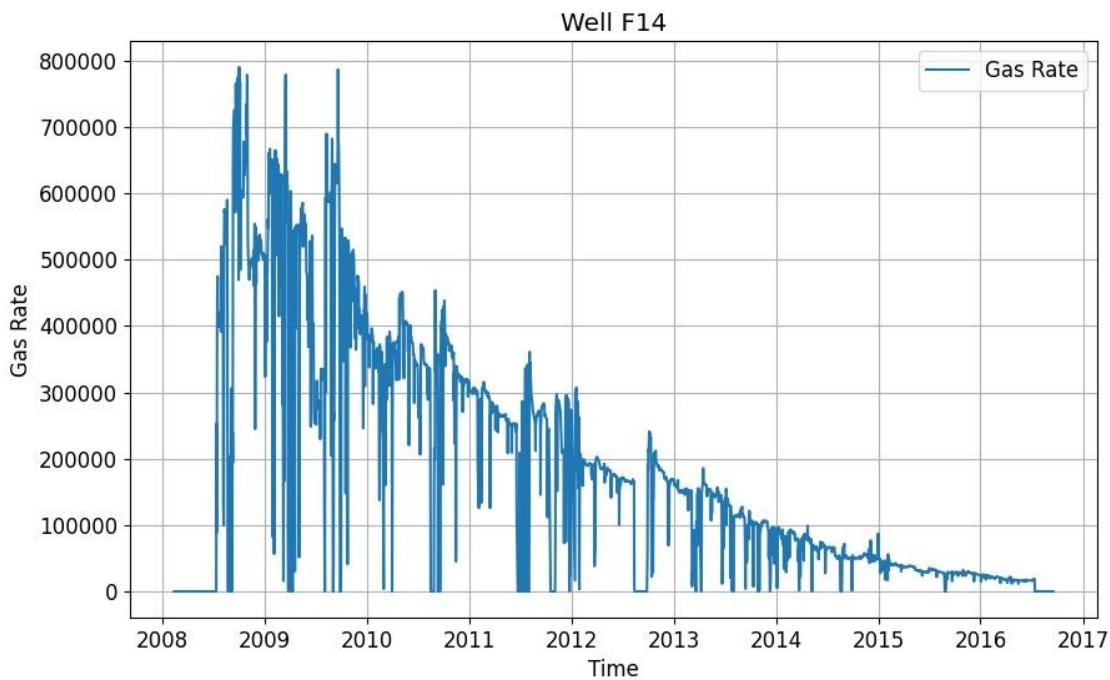
	BTHP	BTHT	DP	CS	WHP	WHT	OIL	GAS	WATER
count	2365,00	2365,00	2365,00	2365,00	2365,00	2365,00	2365,00	2365,00	2365,00
mean	249,09	102,10	209,74	78,48	39,36	87,22	1217,04	180380,41	2968,53
std	14,47	3,35	22,44	24,81	12,61	3,40	1052,87	149356,18	934,56
min	135,63	54,64	103,14	20,30	27,19	48,19	100,33	5928,54	100,16
25%	241,72	99,73	197,39	55,85	31,28	86,25	322,41	49848,46	2694,00
50%	247,56	101,14	208,75	97,92	32,84	87,74	950,38	151088,07	3203,30
75%	261,78	105,05	230,24	100,00	42,23	88,86	1876,97	285954,06	3541,83
max	281,30	106,77	239,84	100,00	115,06	93,51	5644,37	786328,36	5691,77



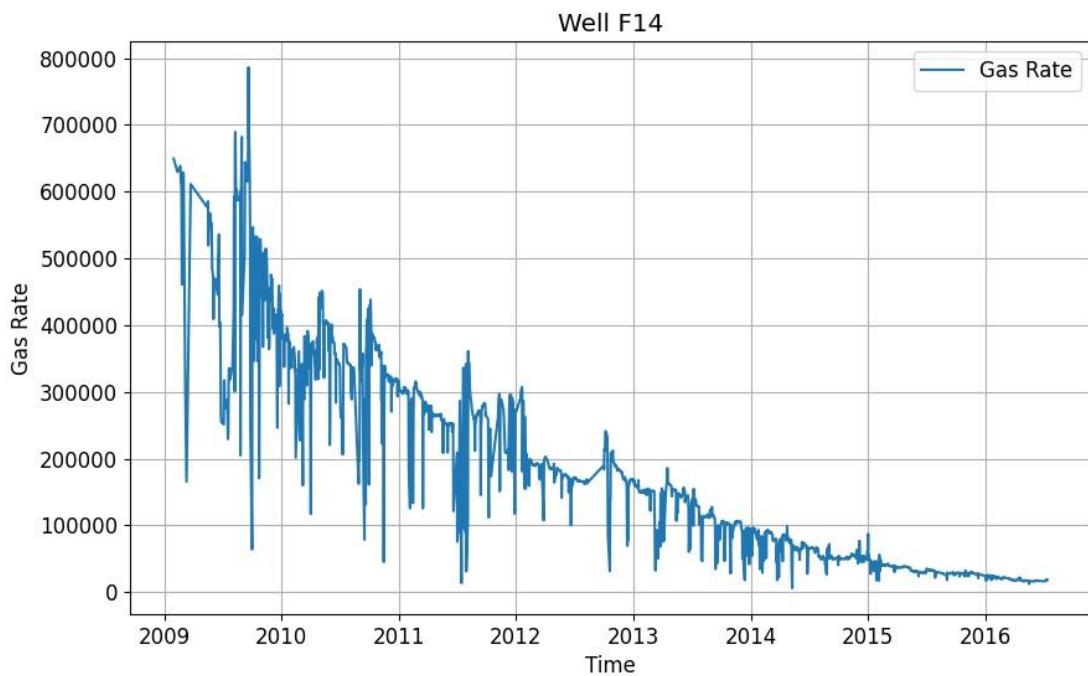
Hình 3-6: Biểu đồ lưu lượng dầu khai thác trước khi dữ liệu được xử lý



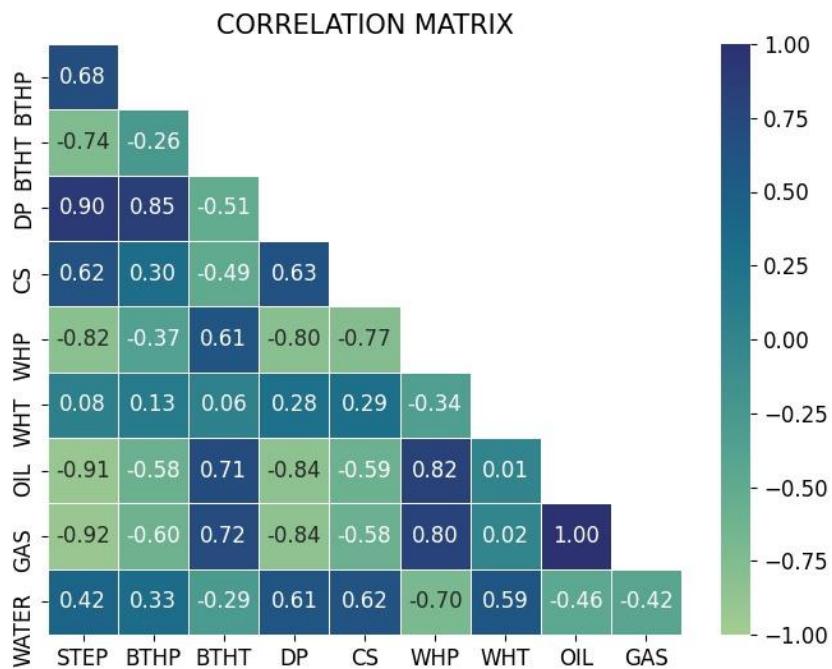
Hình 3-7: Biểu đồ lưu lượng khí khai thác trước khi dữ liệu được xử lý



Hình 3-8: Biểu đồ lưu lượng khí khai thác trước khi dữ liệu được xử lý



Hình 3-9: Biểu đồ lưu lượng khí khai thác sau khi dữ liệu được xử lý



Hình 3-10: Ma trận tương quan giữa các đặc trưng

Bảng 3-3: Dữ liệu của giếng F14, mỏ Volve sau khi được tiền xử lý

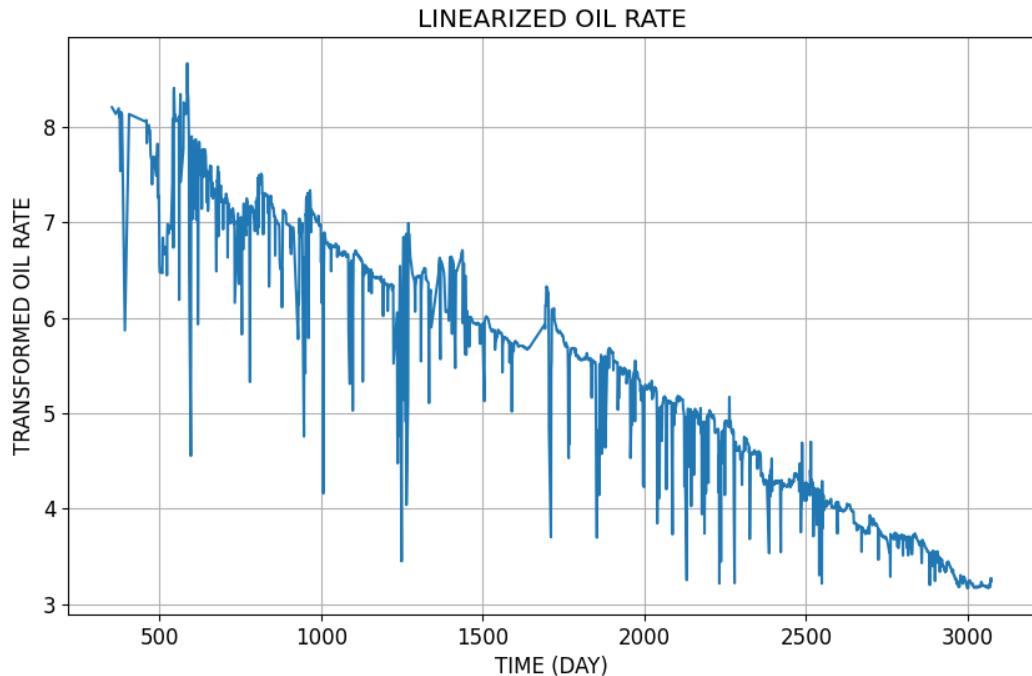
NO	DATE	BTHP	BTHT	DP	CS	WHP	WHT	OIL	GAS	WATER
0	30/01/2009	257,44	105,34	163,29	35,30	94,15	73,62	4535,43	649388,07	298,19
1	11/02/2009	261,48	105,36	164,35	34,70	97,13	80,24	4379,88	629307,34	143,54
2	20/02/2009	264,39	105,41	166,21	34,78	98,17	78,44	4509,07	638750,17	108,74
3	22/02/2009	266,71	105,40	166,27	34,05	100,44	80,12	4319,02	612912,62	106,60
4	23/02/2009	266,67	105,41	166,51	34,40	100,15	81,01	4417,66	625514,01	117,37
5	25/02/2009	269,71	105,36	166,93	31,05	102,78	76,54	3226,61	460948,01	118,99
6	26/02/2009	268,34	105,43	167,28	34,31	101,06	80,42	4411,90	628668,27	134,19
7	27/02/2009	268,75	105,44	167,40	34,31	101,35	81,08	4376,91	625510,25	152,76
8	28/02/2009	269,14	105,44	167,85	34,28	101,29	78,12	4417,91	626562,21	106,72
9	01/03/2009	269,56	105,45	167,93	34,32	101,63	79,10	4396,74	628354,12	155,97
10	02/03/2009	270,00	105,46	168,02	34,24	101,98	80,54	4381,09	623678,16	163,39
11	07/03/2009	281,30	105,22	166,24	24,94	115,06	73,90	2208,96	316638,28	104,01
12	11/03/2009	273,08	105,14	166,64	25,03	106,44	64,78	1185,05	165437,35	174,74
13	24/03/2009	242,68	105,20	159,26	36,77	83,42	79,94	4379,47	611263,11	207,66
...
2350	27/06/2016	269,76	100,17	238,90	45,11	30,86	89,05	101,46	16274,31	2976,42
2351	28/06/2016	269,88	100,17	238,91	45,13	30,97	88,78	102,26	16444,37	2973,66
2352	29/06/2016	269,83	100,19	238,94	45,20	30,89	88,38	102,38	16458,56	2980,12
2353	30/06/2016	269,83	100,20	238,89	45,16	30,93	88,84	101,47	16397,18	2985,45
2354	01/07/2016	269,89	100,20	238,91	45,16	30,98	88,83	102,89	16369,86	2979,13
2355	02/07/2016	269,79	100,22	238,87	45,18	30,93	89,18	101,58	16365,91	2992,00
2356	03/07/2016	269,79	100,23	238,83	45,11	30,96	89,14	102,43	16269,74	2976,87
2357	04/07/2016	269,78	100,24	238,82	45,08	30,96	89,48	100,67	16263,02	2990,10
2358	05/07/2016	269,77	100,25	238,80	45,08	30,97	89,52	101,88	16284,48	2974,30
2359	07/07/2016	266,20	100,31	238,44	93,42	27,76	89,44	106,19	17427,78	3172,96
2360	08/07/2016	266,04	100,33	238,47	100,00	27,56	89,31	106,30	17541,20	3187,95
2361	09/07/2016	268,81	100,30	239,08	82,19	29,73	87,86	102,09	16681,29	2326,24
2362	10/07/2016	265,92	100,34	238,40	100,00	27,52	89,15	113,38	18753,12	3185,47
2363	11/07/2016	267,77	100,32	238,64	91,16	29,13	89,07	108,84	17979,28	3056,29
2364	12/07/2016	266,00	100,35	238,27	100,00	27,73	89,50	113,84	18543,76	3148,91

3.5 Xây dựng mô hình học máy hồi quy

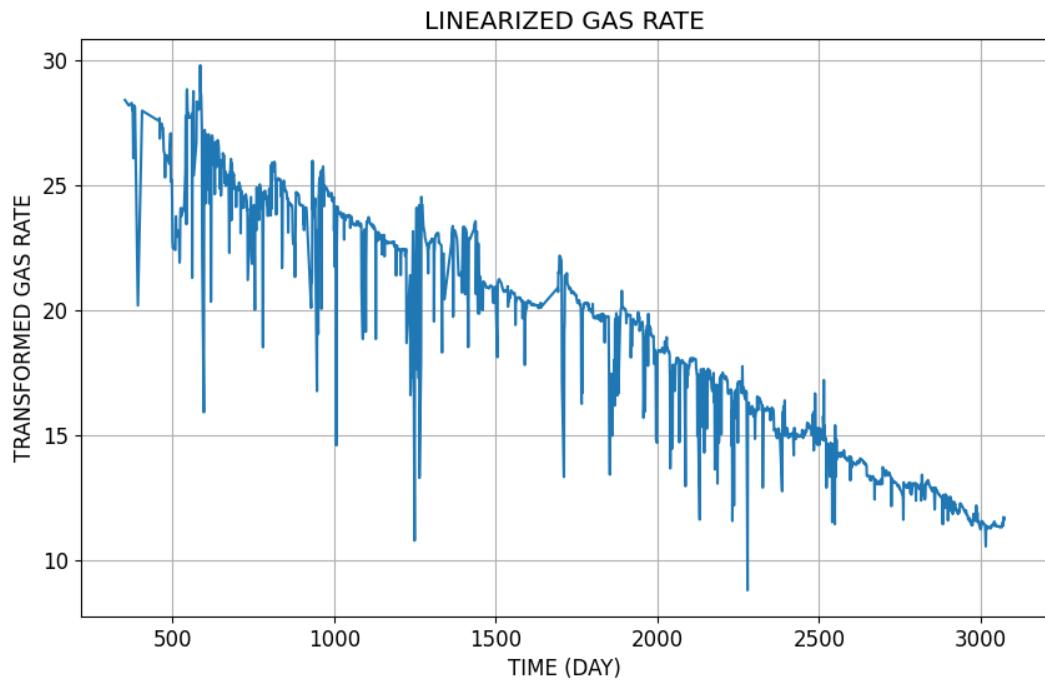
Ta nhận xét rằng, thông số dự báo (OIL hoặc GAS) có dạng phi tuyến tính theo thời gian. Mặt khác, các thông số lưu lượng dầu khai thác là một dạng dữ liệu theo thời gian (time – series) và việc dự báo sản lượng dầu khai cũng phụ thuộc vào biến thời gian. Do đó ta cần tuyến tính hóa các nhãn OIL và GAS sao cho phù hợp với các mô hình học máy hồi quy. Bằng phương pháp thử và sai, ta chấp nhận rằng giá trị lưu lượng của dầu và khí tuân thủ theo quy luật hàm mũ, mà cụ thể là mũ 4. Do đó, để tuyến tính hóa chúng, ta chuyển đổi theo công thức:

$$\tilde{y} = y^{1/4} \quad (3-1)$$

Cũng với phương pháp thử và sai, ta chọn những thông số sau làm đặc trưng đưa vào mô hình hồi quy tuyến tính: STEP, BTHP, DP, WHP. Đối với mô hình hồi quy véc – to hỗ trợ, ta cũng chọn những thông số trên và thêm thông số CS. Trong đó, STEP là biến thời gian, BTHP là áp suất đáy giếng, DP là chênh áp trong ống chống khai thác, WHP là áp suất đầu giếng và CS là độ mở theo phần trăm của ống góp.



Hình 3-11: Lưu lượng dầu khai thác sau khi được tuyến tính hóa



Hình 3-12: Lưu lượng khí khai thác sau khi được tuyến tính hóa

Để xây dựng được mô hình học máy hồi quy tuyến tính, ta cần các thư viện được hỗ trợ cho python sau: numpy, pandas, matplotlib, sklearn. Quy trình xây dựng mô hình học máy nói chung gồm các bước chính:

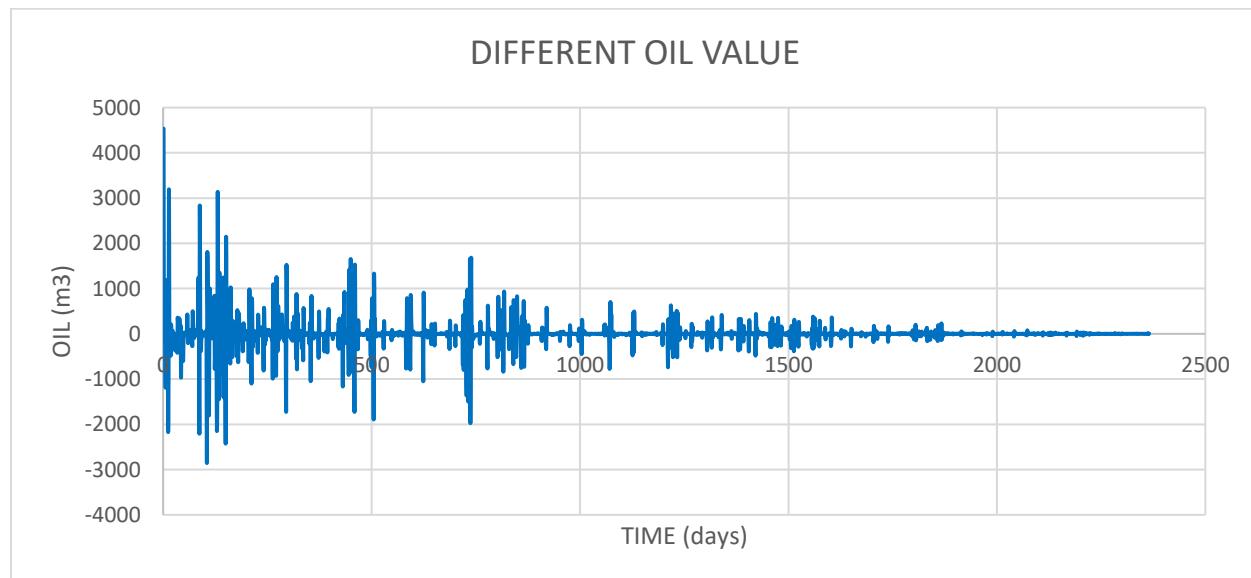
- 1) Xác định đặc trưng và nhãn từ bộ dữ liệu;
- 2) Chia đặc trưng và nhãn thành 3 phần huấn luyện, xác thực và kiểm tra tương ứng;
- 3) Huấn luyện mô hình;
- 4) Xác thực mô hình, nếu kết quả xác thực không đủ tốt, quay lại bước lựa chọn đặc trưng hoặc/ và thay đổi chiến lược huấn luyện mô hình đến khi kết quả xác thực nằm trong các ràng buộc cho phép;
- 5) Dự báo trên tập kiểm tra; đánh giá sai số, hệ số xác định của mô hình và trực quan hóa dữ liệu.

3.6 Xây dựng mô hình mạng nơ – ron hồi quy

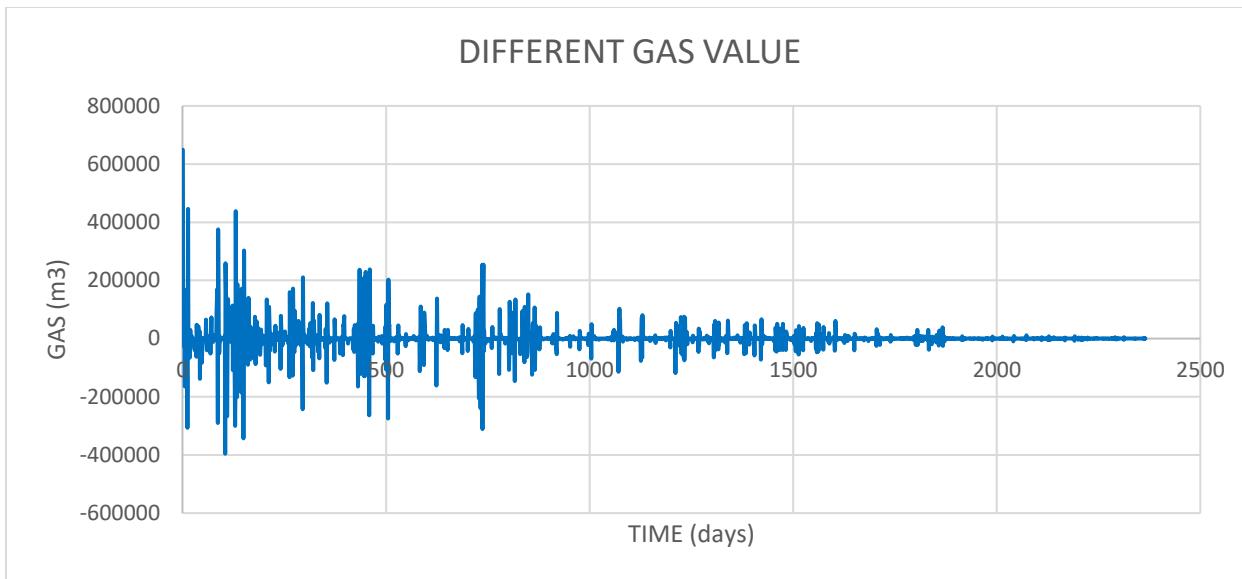
Với bộ dữ liệu có thứ tự hoặc dạng time – series, họ mô hình học sâu mạng nơ – ron hồi quy RNN có nhiều tiềm năng trong việc dự báo xu hướng của các giá trị mới trong tương lai. Với bộ dữ liệu mà ta có, nếu ta chọn biến “OIL” để xây dựng mô hình, trong đó biến “OIL” vừa đóng vai trò là đặc trưng, vừa đóng vai trò là nhãn. Trường hợp ta dự báo sản lượng khí khai thác, “GAS” cũng đóng vai trò tương tự như “OIL”.

Các bước hiện thực cụ thể như sau:

- I) Đây là một mô hình khá phức tạp nên đầu tiên, ta chuẩn bị các hàm hỗ trợ để việc hiện thực dễ dàng hơn và rút ngắn thời gian xử lý các công việc giống nhau lặp đi lặp lại nhiều lần. Các hàm này bao gồm:
 - ***difference***: Dùng để tính chênh lệch giữa hai giá trị liền kề trong tập dữ liệu đầu vào. Mục đích là ta sẽ dự đoán xu hướng tăng giảm của dữ liệu mà không trực tiếp dự báo trên các giá trị thực của chúng nhằm làm tăng tính ổn định của mô hình. Ta thấy rằng, trong hình 3-13 và 3-14, các điểm dữ liệu dường như chỉ dao động quanh đường thẳng $y = 0$. Do đó, việc áp dụng chiến lược này sẽ giúp mô hình tăng đáng kể mức độ chính xác khi dự báo.



Hình 3-13: Lưu lượng dầu sau khi gọi hàm



Hình 3-14: Lưu lượng khí sau khi gọi hàm

- ***invert_difference***: Hàm này trả lại các giá trị ban đầu cho bộ dữ liệu trước khi gọi hàm ***difference*** nhằm tìm ra các giá trị thực ở đầu ra của mô hình và đánh giá chúng so với dữ liệu thực tế.
- ***build_dataset***: Để mô hình dự đoán được xu hướng kế tiếp với một cơ sở các giá trị đầu vào, ta cần chia tập dữ liệu đầu vào thành nhiều mẫu nhỏ. Ví dụ ta có bộ dữ liệu [10, 20, 30, 40, 50, 60, 70, 80, 90], thì cứ 5 giá trị được đưa vào mô hình thì mô hình sẽ phải dự báo giá trị thứ 6. Sau khi gọi hàm, bộ dữ liệu sẽ trở thành:

$$\begin{bmatrix} [10, 20, 30, 40, 50][60] \\ [20, 30, 40, 50, 60][70] \\ [30, 40, 50, 60, 70][80] \\ [40, 50, 60, 70, 80][90] \end{bmatrix}$$

Khi đó ta có tập các véc – tơ đặc trưng là một ma trận 3 chiều $1 \times 5 \times n$, trong đó n là số véc – tơ đặc trưng (hay số mẫu):

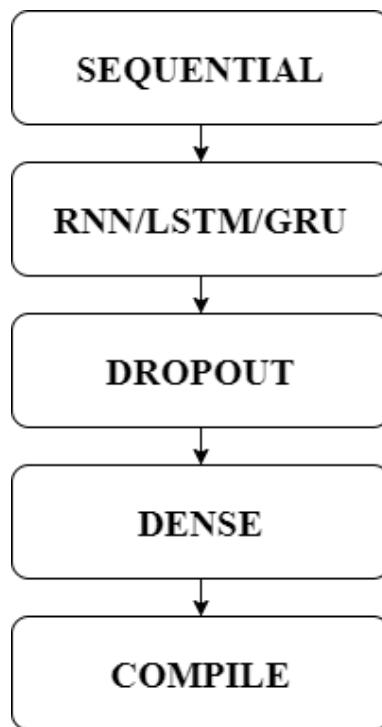
$$\begin{bmatrix} [10, 20, 30, 40, 50] \\ [20, 30, 40, 50, 60] \\ [30, 40, 50, 60, 70] \\ [40, 50, 60, 70, 80] \end{bmatrix}$$

Bên cạnh đó, ta cũng có tập các nhãn là một ma trận 3 chiều $1 \times 1 \times n$:

$$\begin{bmatrix} [60] \\ [70] \\ [80] \\ [90] \end{bmatrix}$$

- ***split_dataset***: Hàm này dùng để tách bộ dữ liệu gốc thành các tập huấn luyện, xác thực và thử nghiệm
- ***scale***: Dùng để chuyển giá trị của bộ dữ liệu về khoảng [-1, 1] (MinMaxScaler) nhằm làm tăng tốc độ hội tụ của mô hình.
- ***invert_scale***: tương ứng với scale, ta có *invert_scale*. Hàm này dùng để giải scale cho bộ dữ liệu.
- ***training***: Hàm *training* dùng để xây dựng và huấn luyện mô hình.
- ***predict***: Hàm *predict* là hàm dự báo xu hướng tiếp theo của một bộ số liệu đưa vào hàm.
- ***errorMeasure***: Hàm này dùng để tính toán các sai số giữa các giá trị dự báo và giá trị thực.
- ***graph***: Hàm *graph* dùng để trực quan hóa các tập dữ liệu.

- **assemble:** Hàm assemble đóng vai trò như “kịch bản chính” của toàn bộ công việc mà ta phải thực hiện. Trong hàm này, ta tiến hành gọi các hàm bên trên theo thứ tự mà được định sẵn để hiện thực giải thuật của ta.
 - **run:** Mục đích của hàm này là tiến hành khởi chạy hàm assemble sau khi load bộ dữ liệu vào chương trình.
- 2) Xây dựng mô hình học sâu qua hàm **training**: Mô hình học sâu mạng nơ – ron hồi quy được xây dựng dựa trên các lớp Sequential, SimpleRNN/GRU/LSTM, Dropout, Dense và Compile. Lớp Sequential được thêm vào nhằm đảm bảo thứ tự trước sau cho các lớp mà ta thêm vào. Các lớp SimpleRNN, GRU hay LSTM chính là lớp chính của mô hình. Dropout được sử dụng để bỏ bớt một số mẫu ngẫu nhiên trong bộ dữ liệu đưa vào lớp tiếp theo nhằm tránh hiện tượng quá khớp. Lớp Dense được gọi là một lớp “kết nối đầy đủ”, toàn bộ các đơn vị của các layer trước được kết nối với toàn bộ các đơn vị của lớp hiện tại. Hàm Compile được sử dụng để huấn luyện mô hình sử dụng các thuật toán tối ưu hóa như adam, SGD, RMSprop, ... Cụ thể cấu trúc mạng của các mô hình học sâu được xây dựng như sau: Lớp Sequential không có các giá trị đặc tả nào vì như đã nói bên trên, nó là lớp xác định thứ tự của các lớp khác trong mô hình. Các lớp SimpleRNN, GRU hoặc LSTM có 1 đơn vị đầu vào với hàm kích hoạt là sigmoid. Lớp Dropout được truyền vào giá trị 0.3 xác định số điểm dữ liệu ngẫu nhiên bị loại bỏ khỏi tập huấn luyện ban đầu sau mỗi lần huấn luyện. Lớp Dense được truyền vào giá trị 1 xác định số nút đầu ra của mô hình. Lớp Compile sử dụng sai số trung bình bình phương để đánh giá mô hình sau mỗi lần huấn luyện, phương pháp tối ưu hóa mà ta chọn cho họ mô hình mạng nơ – ron hồi quy là adam cũng được truyền vào lớp Compile. Bên cạnh đó, số epoch được xác định cho mô hình là 10 tương ứng với 10 lần huấn luyện, batch size được gán bằng 1 tương ứng lần lượt chỉ có 1 điểm dữ liệu được truyền vào mô hình để huấn luyện. Số điểm dữ liệu được học trước khi mô hình dự đoán điểm dữ liệu tiếp theo là 5, nghĩa là cứ 5 điểm dữ liệu được học mô hình phải đưa ra dự báo giá trị điểm dữ liệu tiếp theo liền sau đó.



Hình 3-15: Cấu trúc họ mạng nơ – ron hồi quy

- 3) Để kết hợp nhịp nhàng giữa các hàm, ta cần phải hiện thực hàm **assemble**. Hàm **assemble** đóng vai trò như kịch bản của toàn bộ chương trình, nó xác định hàm nào được gọi trước, hàm nào được gọi sau sao cho đầu tra của hàm này là đầu vào của hàm kia, từ đó chương trình được thực thi xuyên suốt. Đầu tiên ta phải gọi hàm **difference** cho bộ số liệu đầu vào để tạo thành một bộ dữ liệu về xu hướng biến đổi của các mẫu dữ liệu. Sau đó, ta gọi hàm **build_dataset** để xây dựng bộ dữ liệu phù hợp với yêu cầu của họ mạng nơ – ron hồi quy. Với bộ dữ liệu có được, ta tiến hành phân chia các tập dữ liệu dùng để huấn luyện (70%), xác thực (15%) và kiểm tra (15%) thông qua lời gọi hàm **split_dataset**. Ké đến, ta chuyển khoảng giá trị của các bộ dữ liệu về khoảng [-1, 1] thông qua hàm **scale**. Gọi hàm **training** truyền vào bộ dữ liệu huấn luyện để huấn luyện mô hình. Sau khi huấn luyện xong, ta tiến hành dự báo thông qua hàm **predict**, giải scale qua hàm **invert_scale** và trả về giá trị thực qua hàm **invert_difference** theo thứ tự. Sau đó, tính toán sai số cho từng tập dữ liệu qua

hàm *errorMeasure*, hàm này trả về sai số MSE, MAE và hệ số xác định bình phương để đánh giá mô hình. Cuối cùng ta trực quan hóa dữ liệu thông qua hàm *graph*.

Tương tự như xây dựng mô hình học máy hồi quy tuyến tính, trong phần xây dựng mô hình học sâu ta cũng dùng một số thư viện của python như numpy, pandas, sklearn, matplotlib và bổ sung thêm thư viện keras để phụ vụ cho mô hình học sâu.

3.7 Đánh giá kết quả thực hiện

Để đánh giá mức độ hiệu quả của các mô hình, ta dùng phương pháp đánh giá sai số trung bình bình phương và trung bình trị tuyệt đối, cùng với đó là hệ số xác định R²:

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2 \quad (3-2)$$

$$RMSE(y, \hat{y}) = \sqrt{MSE(y, \hat{y})} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2} \quad (3-3)$$

$$MAE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}| \quad (3-4)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y - \hat{y})^2}{\sum_{i=1}^n (y - \bar{y})^2} \quad (3-5)$$

3.7.1 Đánh giá các mô hình học máy

Chính vì quan hệ tương quan rất chặt giữa lưu lượng khai thác của dầu và khí nên mức độ hiệu quả của mỗi mô hình đối với dầu và khí khá tương đồng với nhau. Quá trình huấn luyện cũng không gây ra hiện tượng quá khớp, bằng chứng là cả hai tập xác thực và kiểm tra cũng đều cho kết quả tốt tương tự như tập huấn luyện.

Các thuật toán học máy đòi hỏi dữ liệu phải được xáo trộn trước khi được dùng để huấn luyện. Do đó tính có thứ tự theo thời gian sẽ bị loại bỏ nếu ta không bổ sung thêm một đặc trưng là biến thời gian (STEP). Đồng thời, bộ dữ liệu của ta cũng biến động rất mạnh và không thuộc dạng tuyến tính, nên ta cần tuyến tính hóa lưu lượng dầu và khí trong các mô hình dự báo tương ứng nhằm làm các đại lượng đó phù hợp hơn với mô hình và tăng hiệu suất dự báo của mô hình tương ứng.

Các giá trị dự báo so với các giá trị thực khi biểu diễn trên đồ thị đường như đều nằm trên đường phân giác thứ nhất.

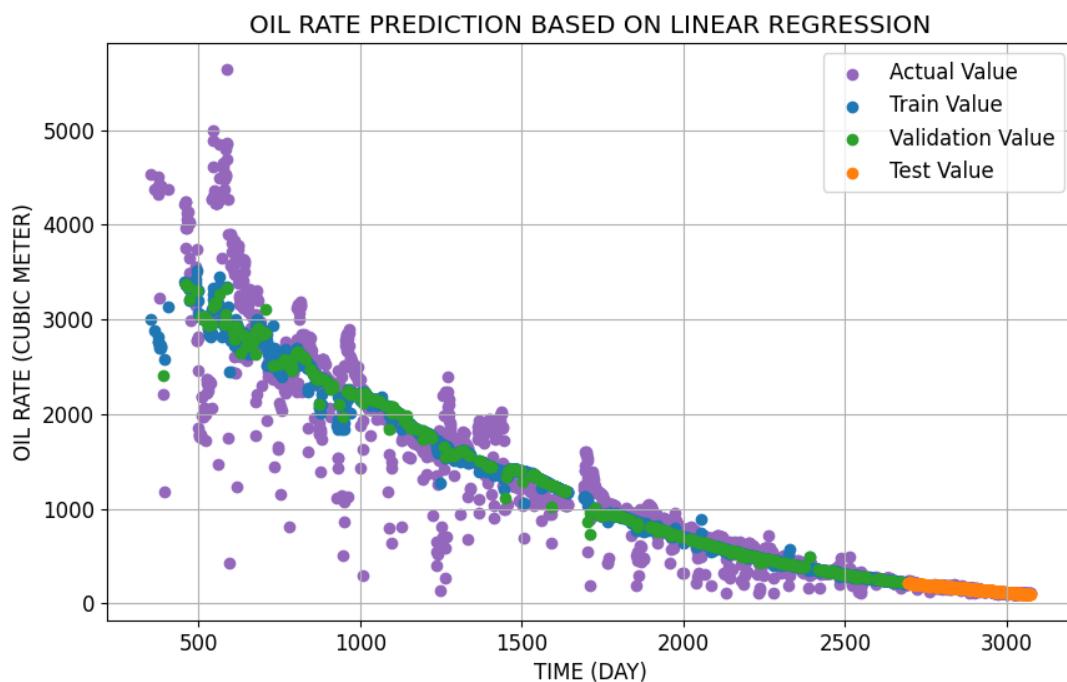
3.7.1.1 Đánh giá mô hình hồi quy tuyến tính (Linear Regression)

Bảng 3-4: Bảng đánh giá mô hình hồi quy tuyến tính dự báo sản lượng dầu khai thác

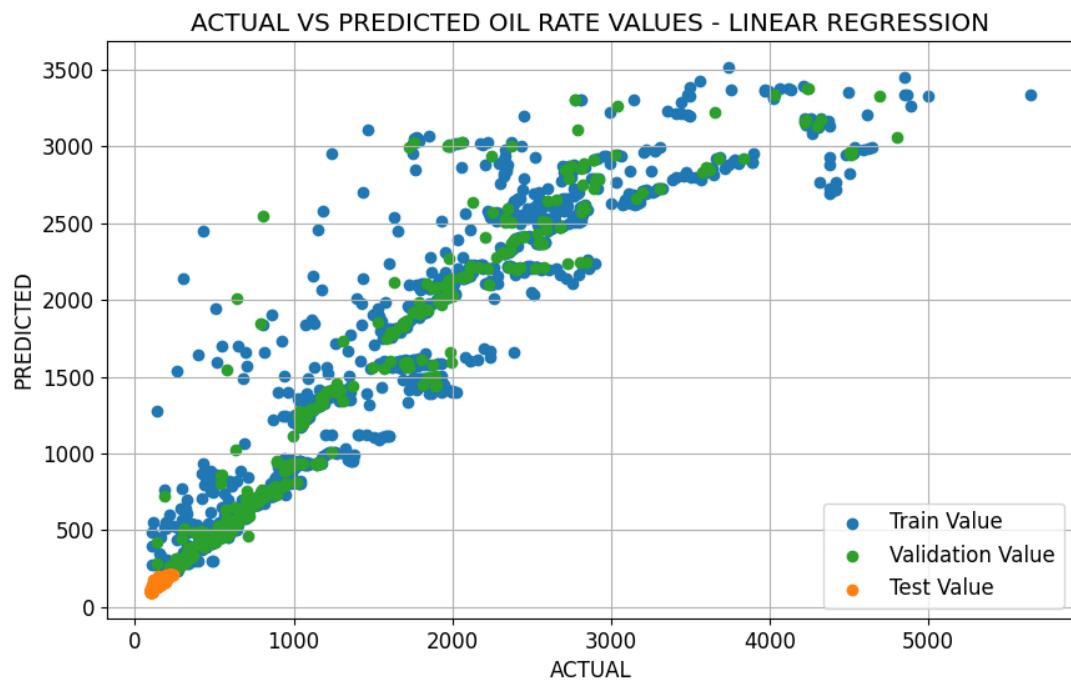
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	372.7573	14.6724	0.9157
Xác thực	366.9816	14.3109	0.9266
Kiểm tra	12.0054	2.9951	0.9070

Bảng 3-5: Bảng đánh giá mô hình hồi quy tuyến tính dự báo sản lượng khí khai thác

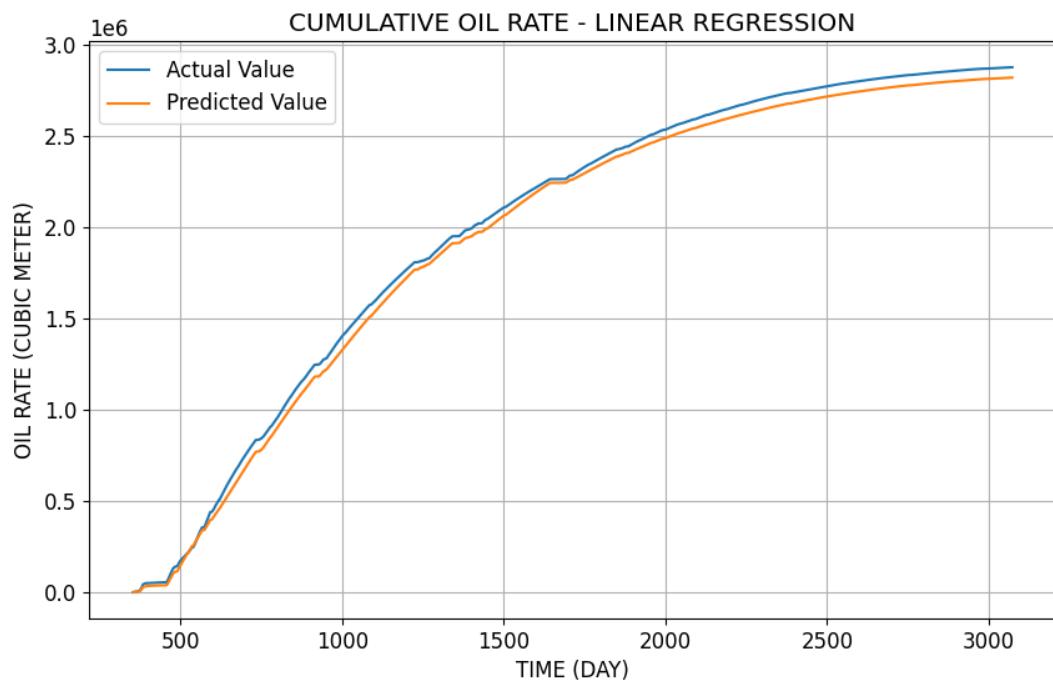
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	52316.4032	174.8368	0.9132
Xác thực	55749.0699	174.8370	0.9189
Kiểm tra	1847.1564	38.1661	0.8750



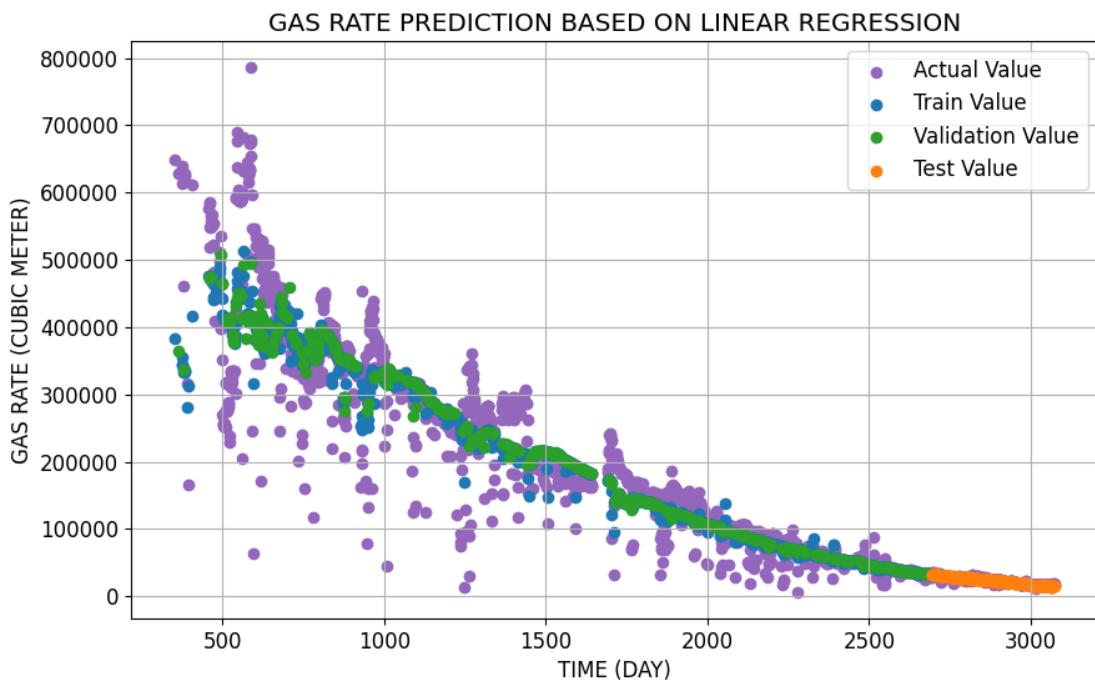
Hình 3-16: Dự báo sản lượng dầu khai thác theo mô hình hồi quy tuyến tính



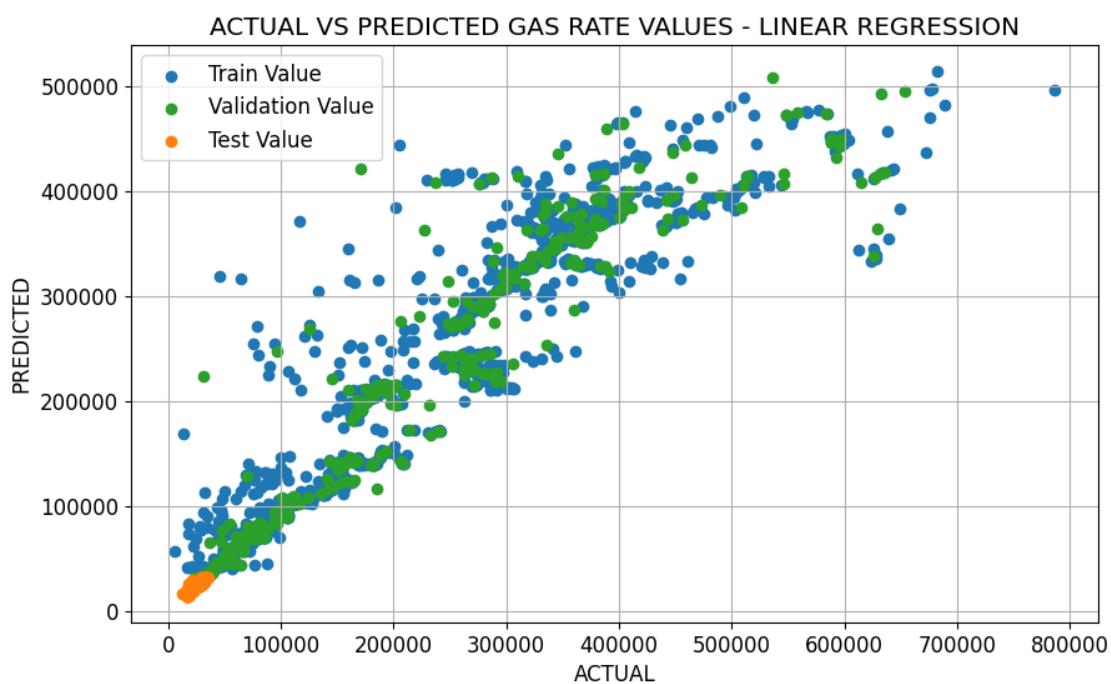
Hình 3-17: Sản lượng dầu khai thác dự báo so với giá trị thực tế - Hồi quy tuyến tính



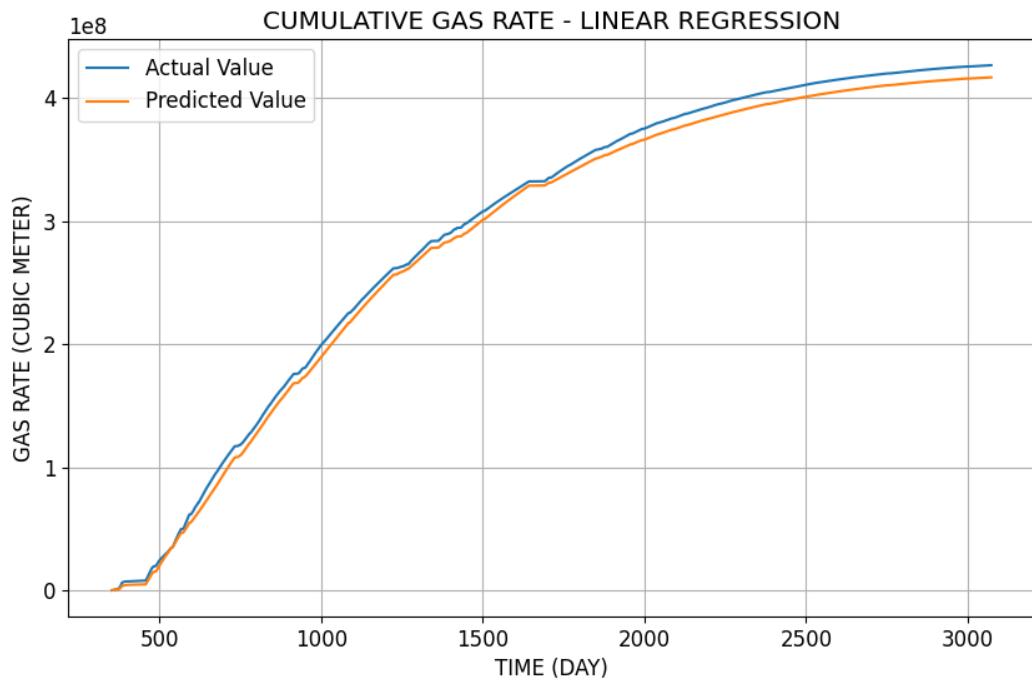
Hình 3-18: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy tuyến tính



Hình 3-19: Dự báo sản lượng khí khai thác theo mô hình hồi quy tuyến tính



Hình 3-20: Sản lượng khí khai thác dự báo so với giá trị thực tế – Hồi quy tuyến tính



Hình 3-21: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy tuyến tính

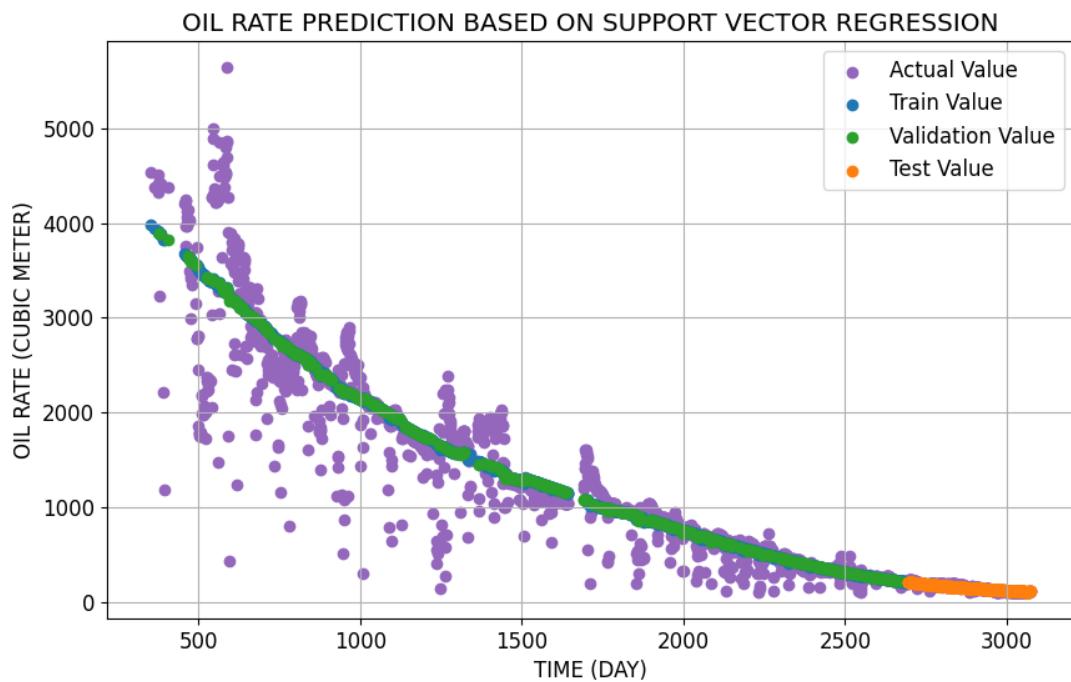
3.7.1.2 Đánh giá mô hình hồi quy véc – tơ hỗ trợ (Support Vector Regression)

Bảng 3-6: Bảng đánh giá mô hình hồi quy Ridge dự báo sản lượng dầu khai thác

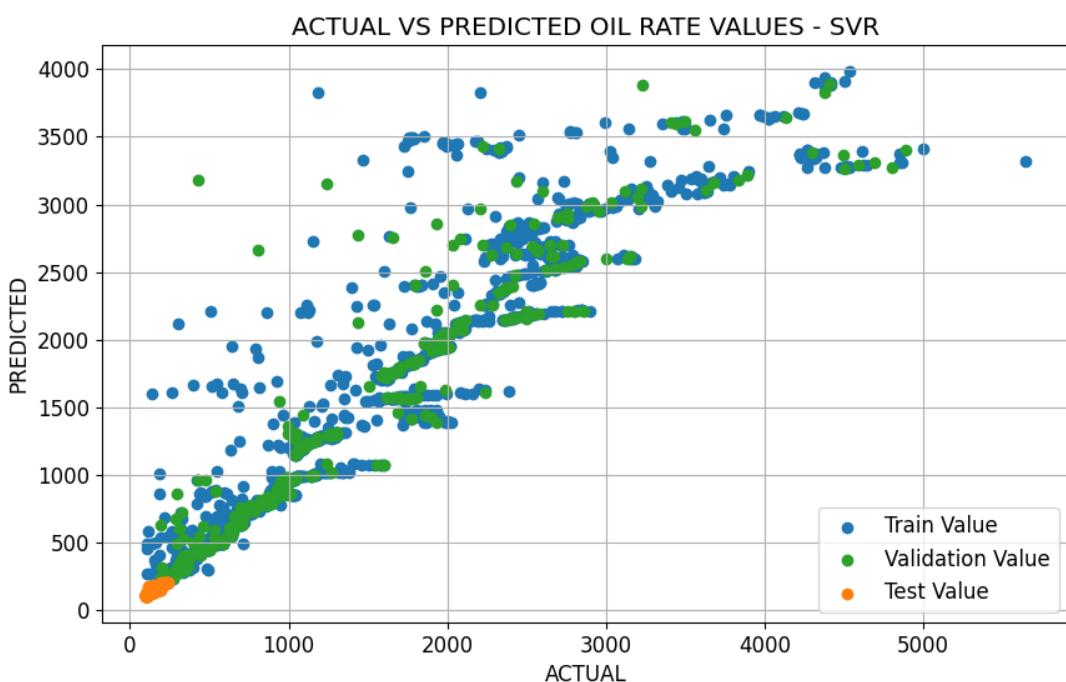
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	375.0697	14.0125	0.9129
Xác thực	373.4134	13.8088	0.9152
Kiểm tra	15.8188	3.5612	0.8444

Bảng 3-7: Bảng đánh giá mô hình hồi quy Ridge dự báo sản lượng khí khai thác

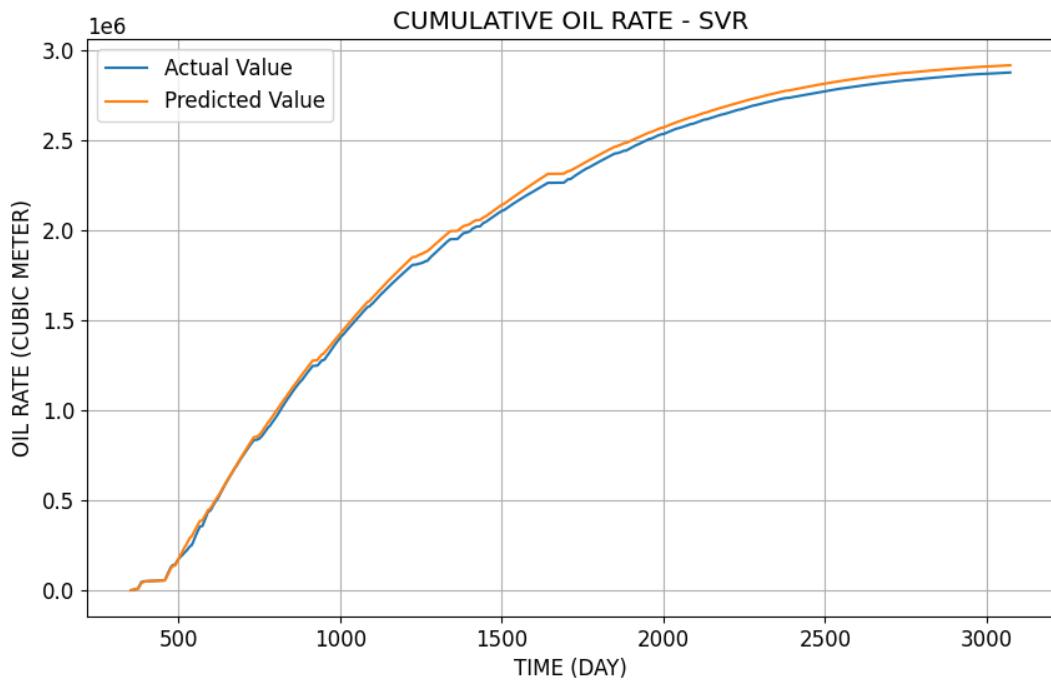
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	53080.5774	170.2241	0.9124
Xác thực	57069.8812	178.5682	0.8870
Kiểm tra	2397.7061	42.7614	0.8132



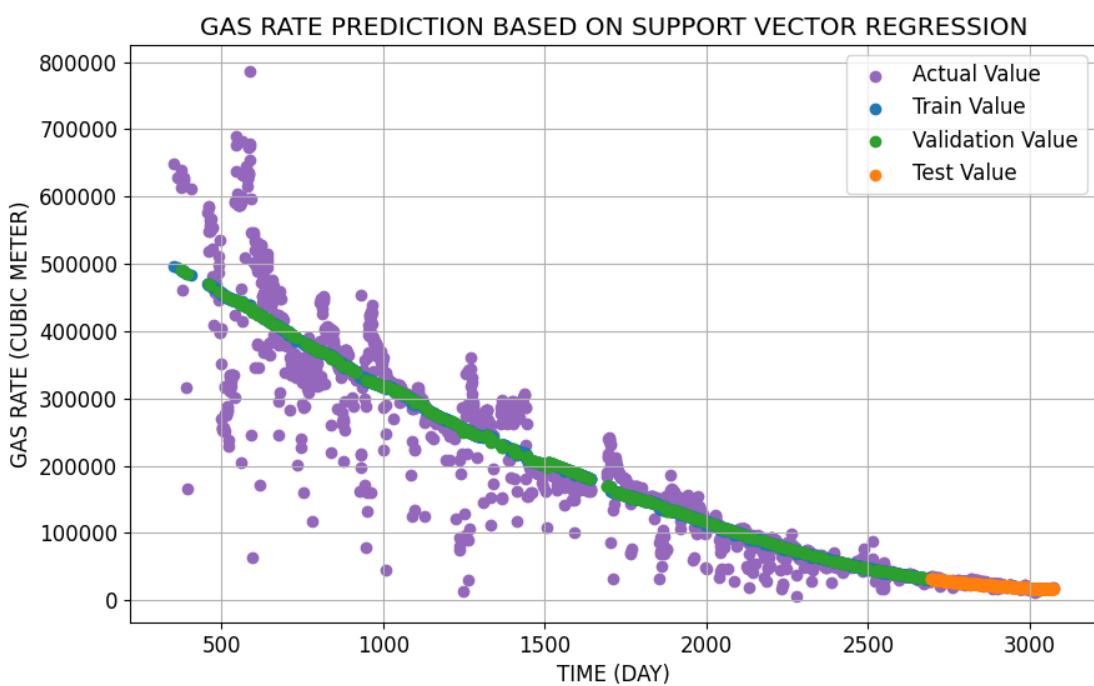
Hình 3-22: Dự báo sản lượng dầu khai thác theo mô hình hồi quy Ridge



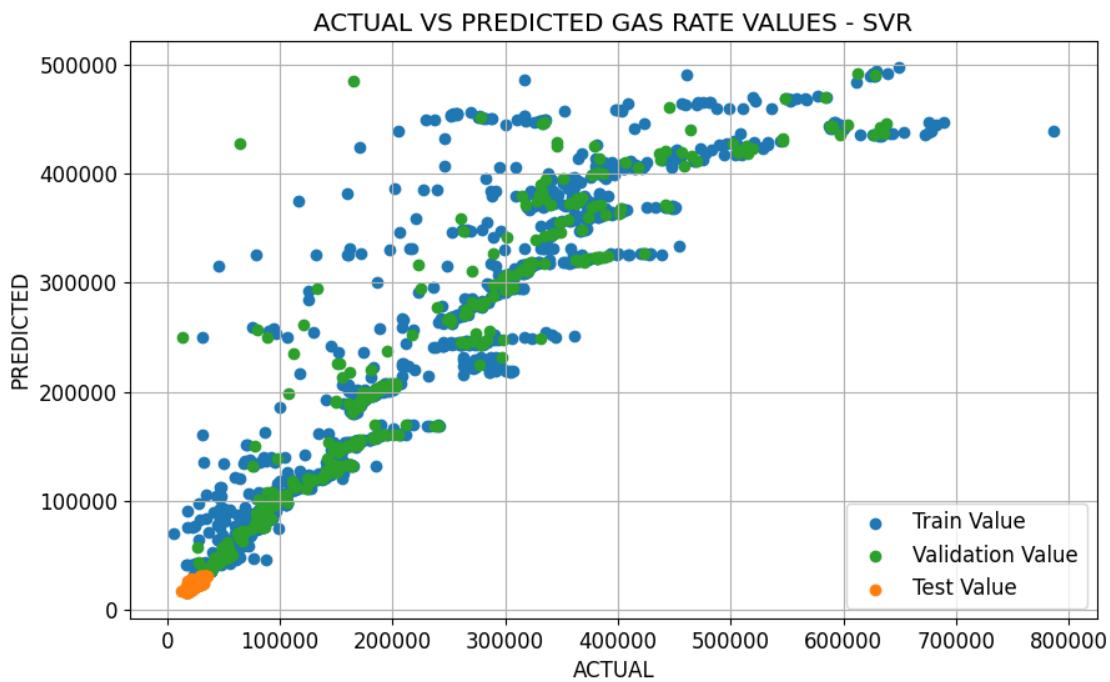
Hình 3-23: Sản lượng dầu khai thác dự báo so với giá trị thực tế – Hồi quy Ridge



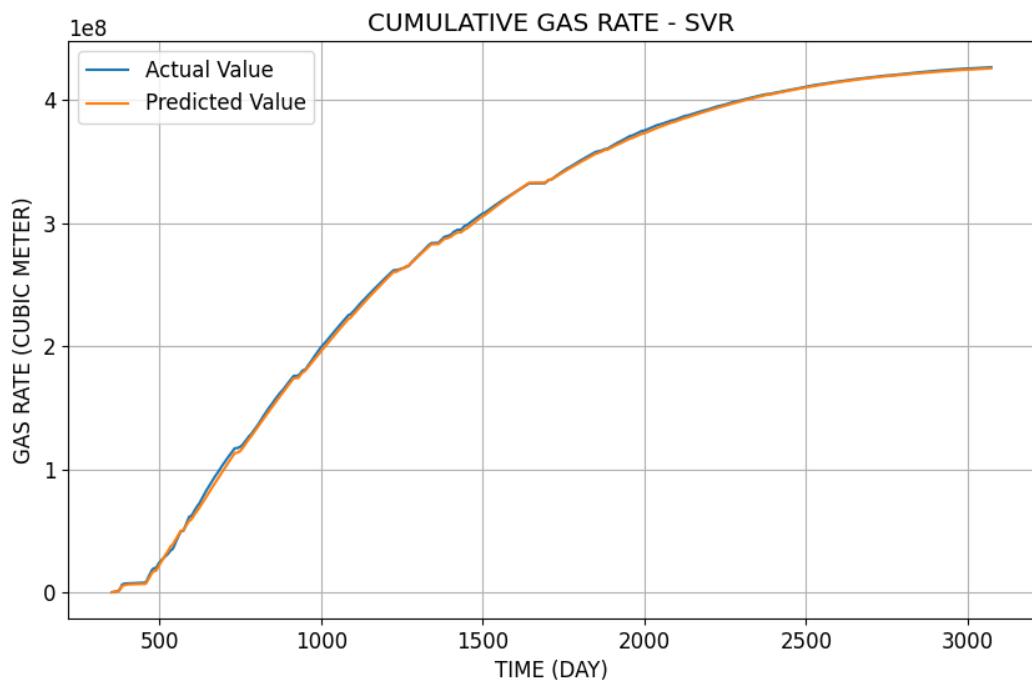
Hình 3-24: Sản lượng dầu khai thác tích lũy theo thời gian – Hồi quy Ridge



Hình 3-25: Dự báo sản lượng khí khai thác theo mô hình hồi quy Ridge



Hình 3-26: Sản lượng khí khai thác dự báo so với giá trị thực tế – Hồi quy Ridge



Hình 3-27: Sản lượng khí khai thác tích lũy theo thời gian – Hồi quy Ridge

3.7.2 Đánh giá các mô hình mạng nơ – ron hồi quy

Khác với các mô hình học máy, các mô hình học sâu chỉ sử dụng một đặc trưng duy nhất để dự báo. Trong nội dung này, để bảo toàn đặc trưng về thứ tự thời gian, ta không xáo trộn tập dữ liệu như trong phần trước. Các kết quả cũng đều cực kỳ tốt như trong các mô hình học máy.

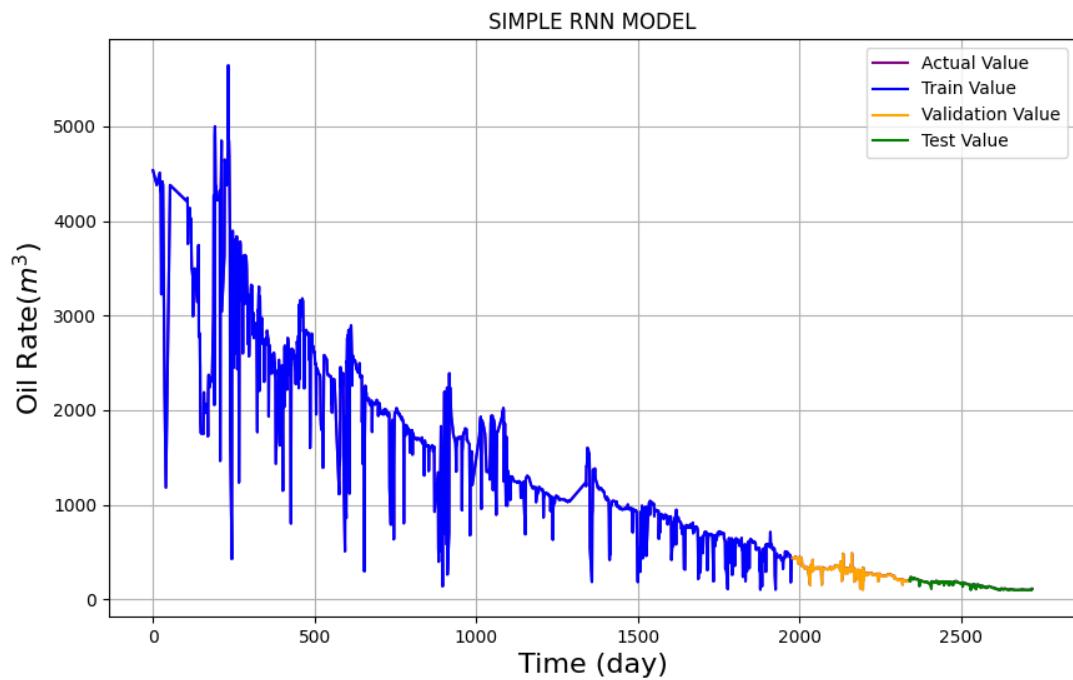
3.7.2.1 Đánh giá mô hình mạng nơ – ron hồi quy đơn giản (*Simple Recurrent Neural Network*)

Bảng 3-8: Bảng đánh giá mô hình mạng nơ – ron hồi quy đơn giản dự báo sản lượng dầu khai thác

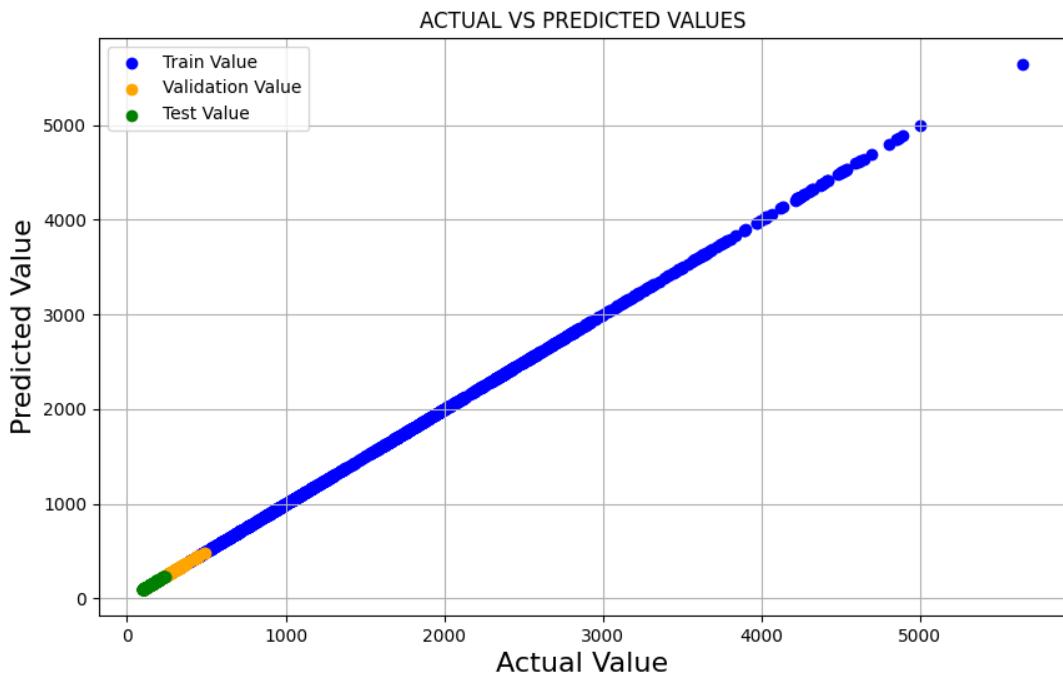
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	2.1230	1.7535	1.0000
Xác thực	2.9087	2.1874	0.9982
Kiểm tra	2.9087	2.1874	0.9942

Bảng 3-9: Bảng đánh giá mô hình mạng nơ – ron hồi quy đơn giản dự báo sản lượng khí khai thác

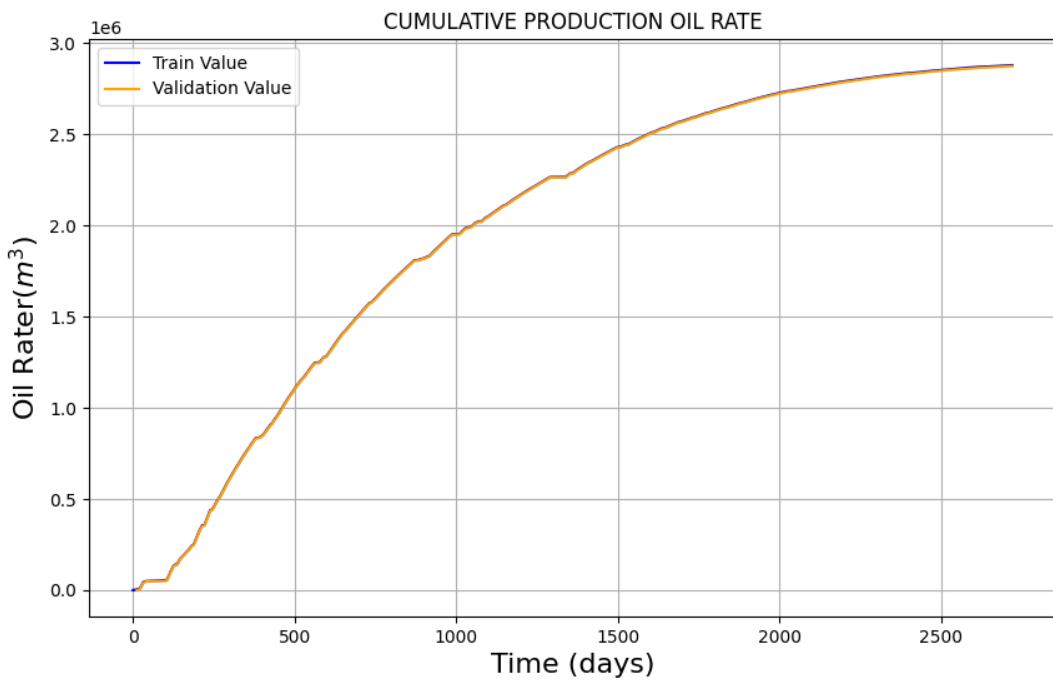
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	203.6990	202.2978	1.0000
Xác thực	206.3152	202.6351	0.9997
Kiểm tra	206.3152	202.6351	0.9985



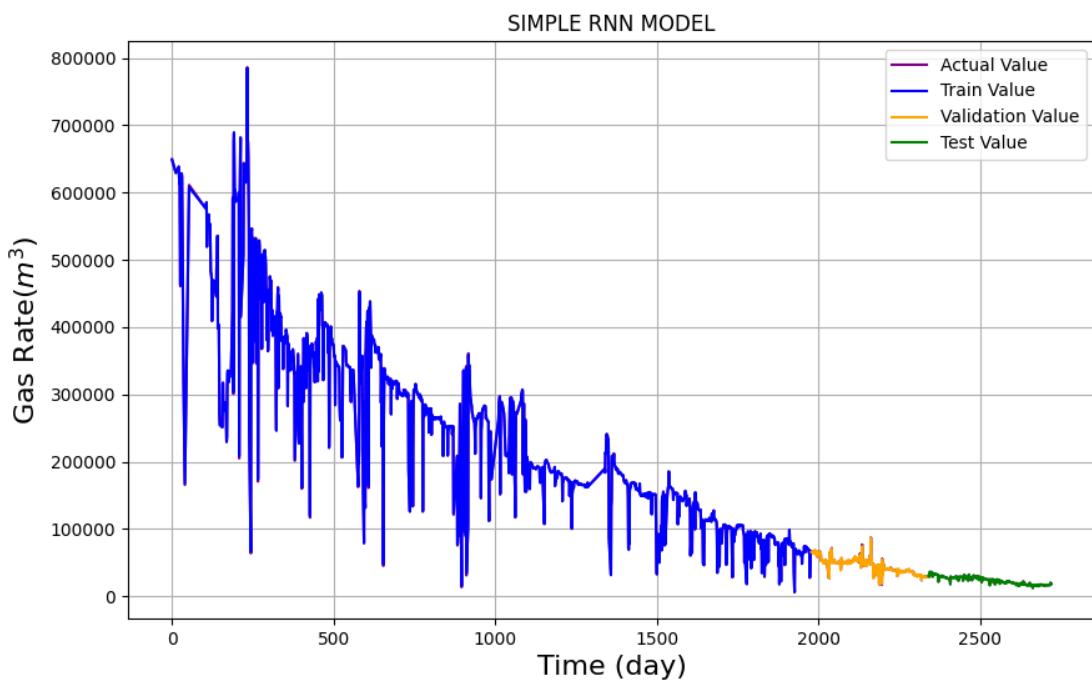
Hình 3-28: Dự báo sản lượng dầu khai thác theo mô hình Simple RNN



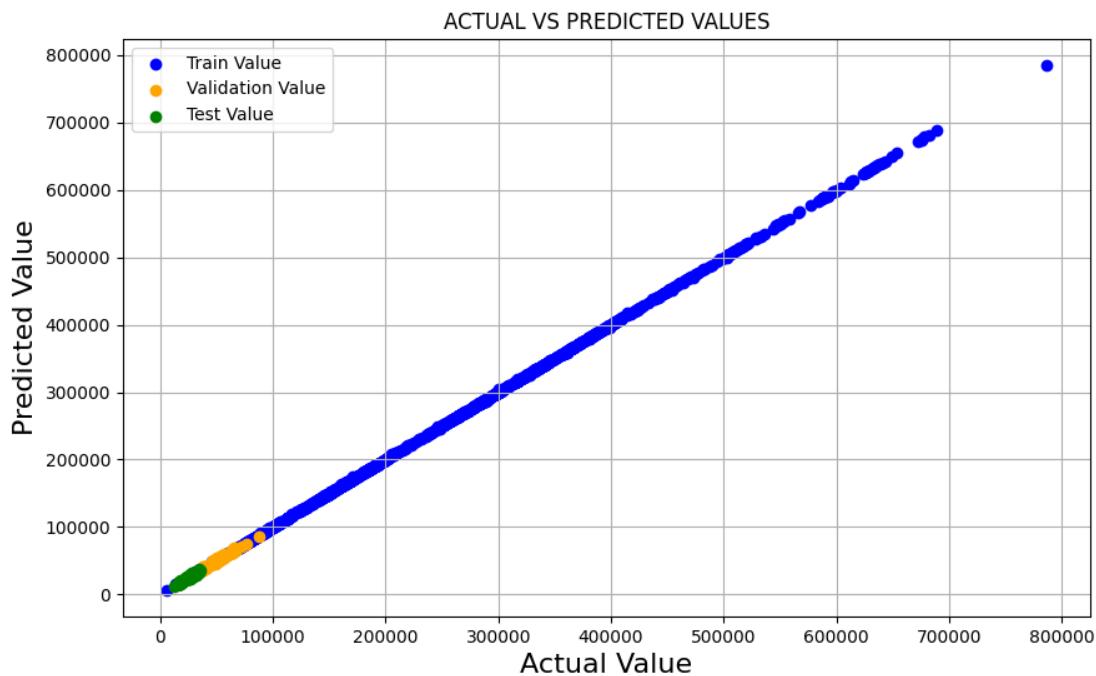
Hình 3-29: Sản lượng dầu khai thác dự báo so với giá trị thực tế – Simple RNN



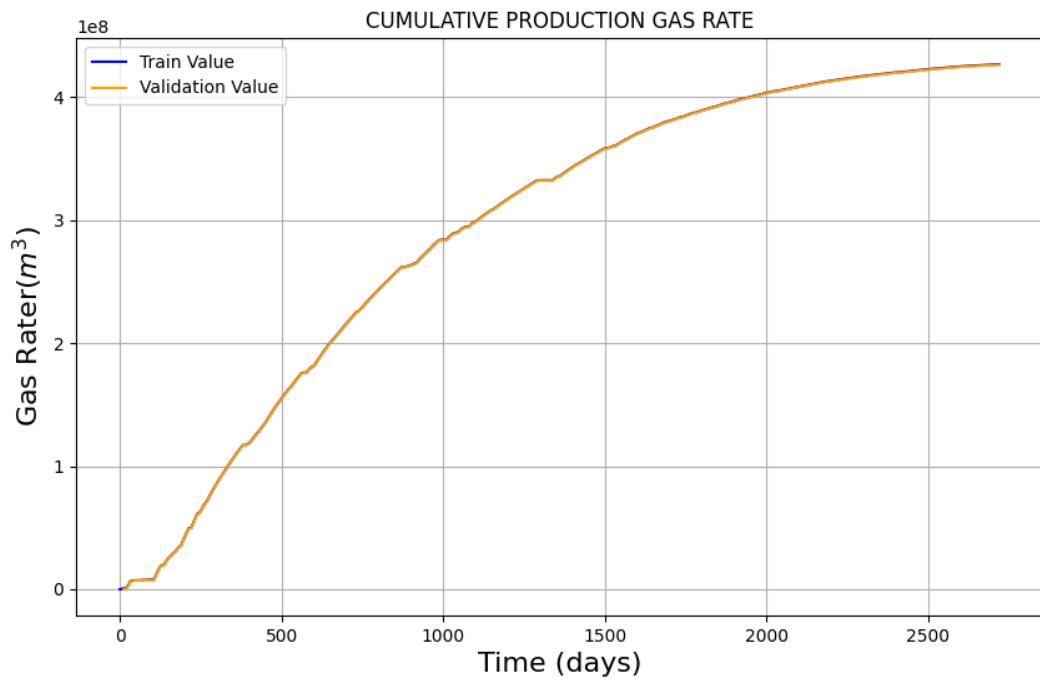
Hình 3-30: Sản lượng dầu khai thác tích lũy theo thời gian – Simple RNN



Hình 3-31: Dự báo sản lượng khí khai thác theo mô hình Simple RNN



Hình 3-32: Sản lượng khí khai thác dự báo so với giá trị thực tế – Simple RNN



Hình 3-33: Sản lượng khí khai thác tích lũy theo thời gian – Simple RNN

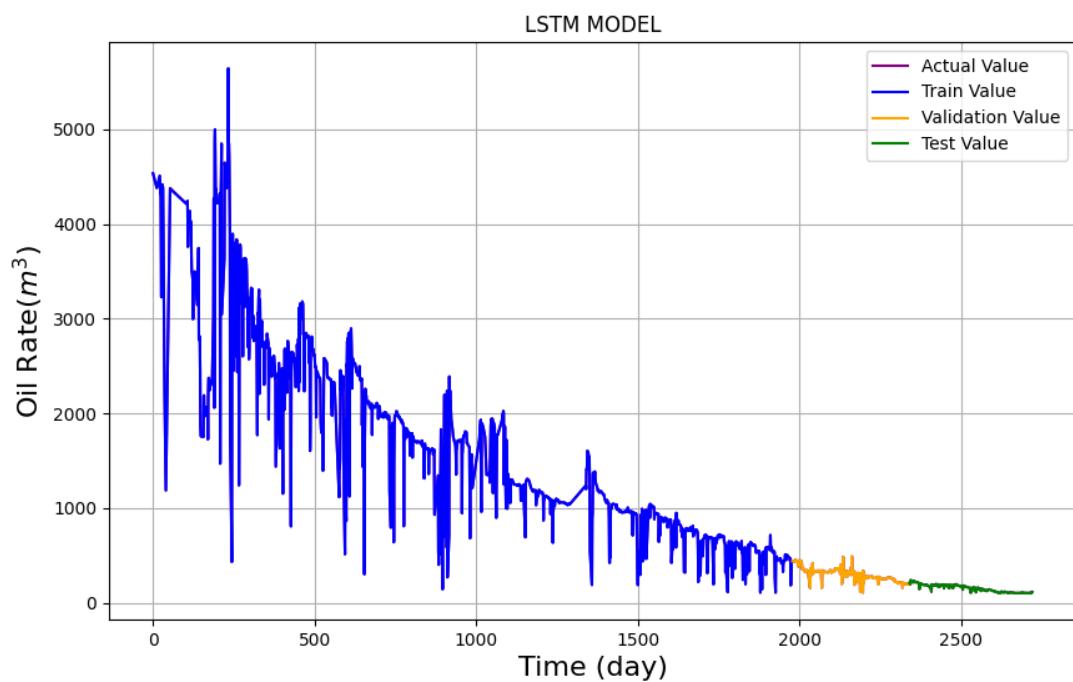
3.7.2.2 Đánh giá mô hình bộ nhớ ngắn hạn dài (Long Short – Term Memory)

Bảng 3-10: Bảng đánh giá mô hình bộ nhớ ngắn hạn dài dự báo sản lượng dầu khai thác

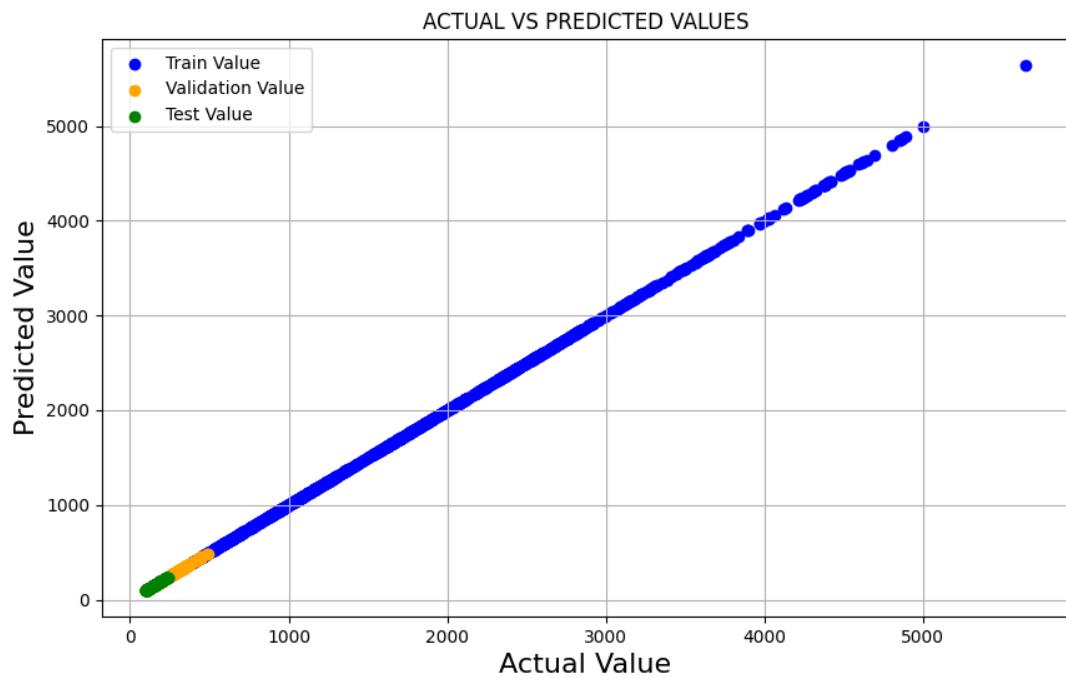
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	5.3721	2.1311	1.0000
Xác thực	0.6896	0.4330	0.9999
Kiểm tra	0.3170	0.2882	0.9999

Bảng 3-11: Bảng đánh giá mô hình bộ nhớ ngắn hạn dài dự báo sản lượng khí khai thác

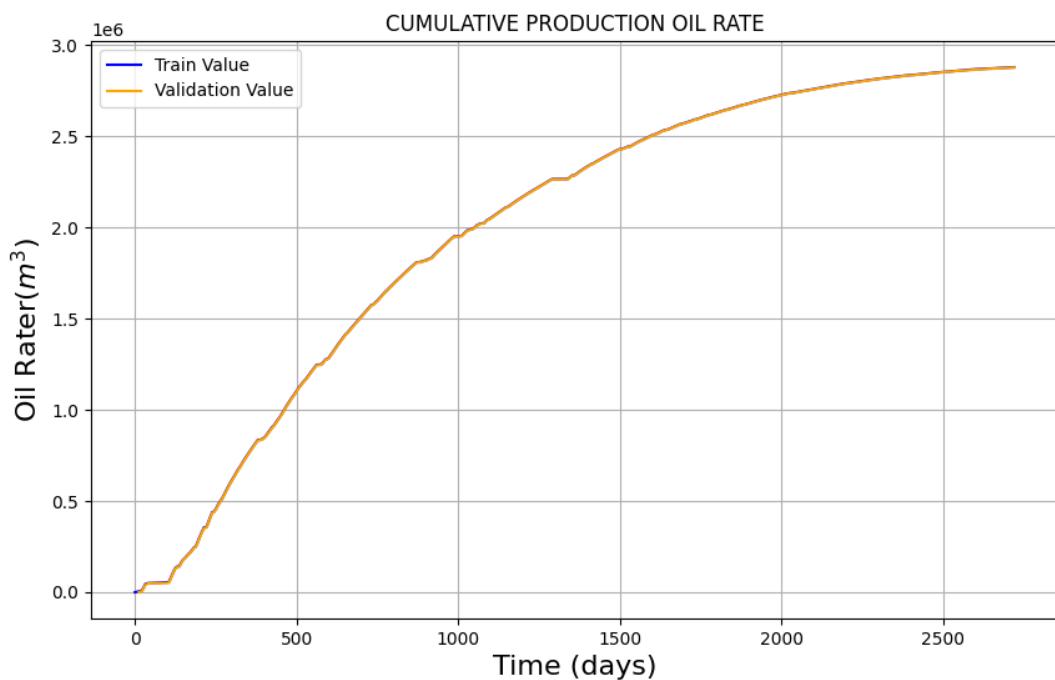
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	111.0014	76.4531	1.0000
Xác thực	59.4249	58.4101	1.0000
Kiểm tra	58.5009	58.4232	0.9999



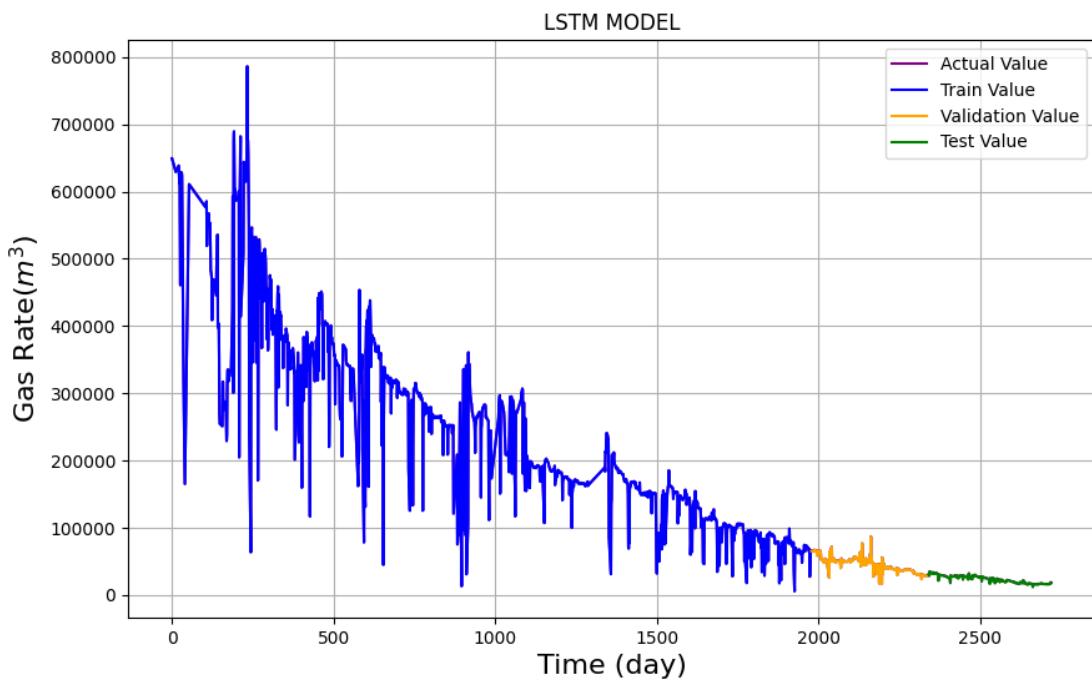
Hình 3-34: Dự báo sản lượng dầu khai thác theo mô hình LSTM



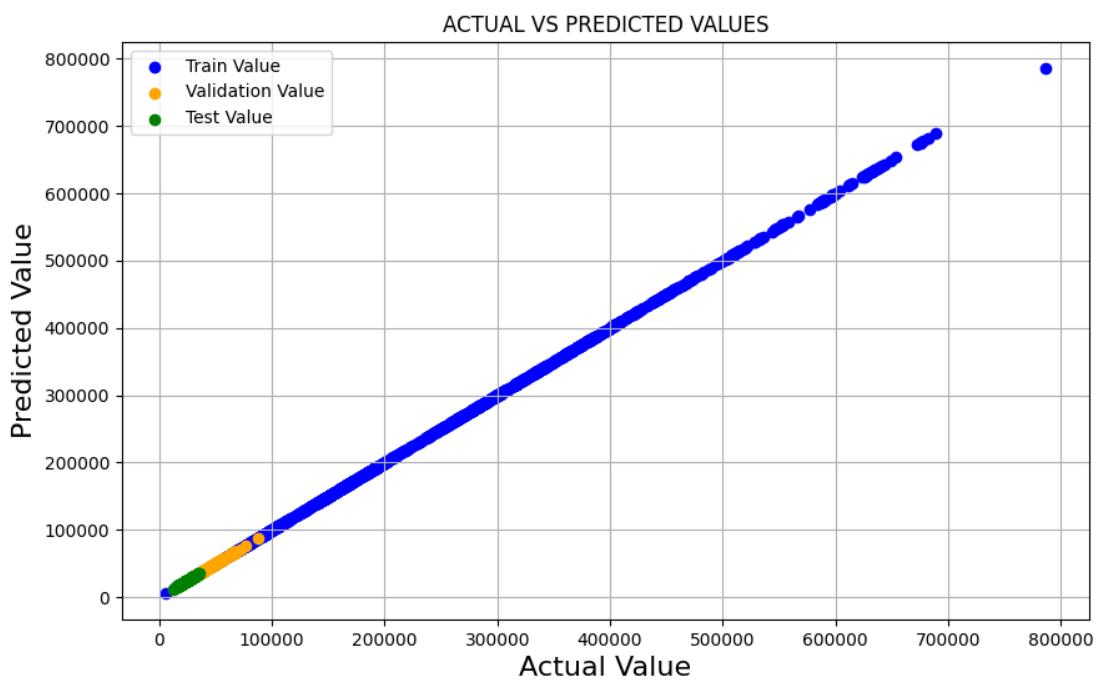
Hình 3-35: Sản lượng dầu khai thác dự báo so với giá trị thực tế – LSTM



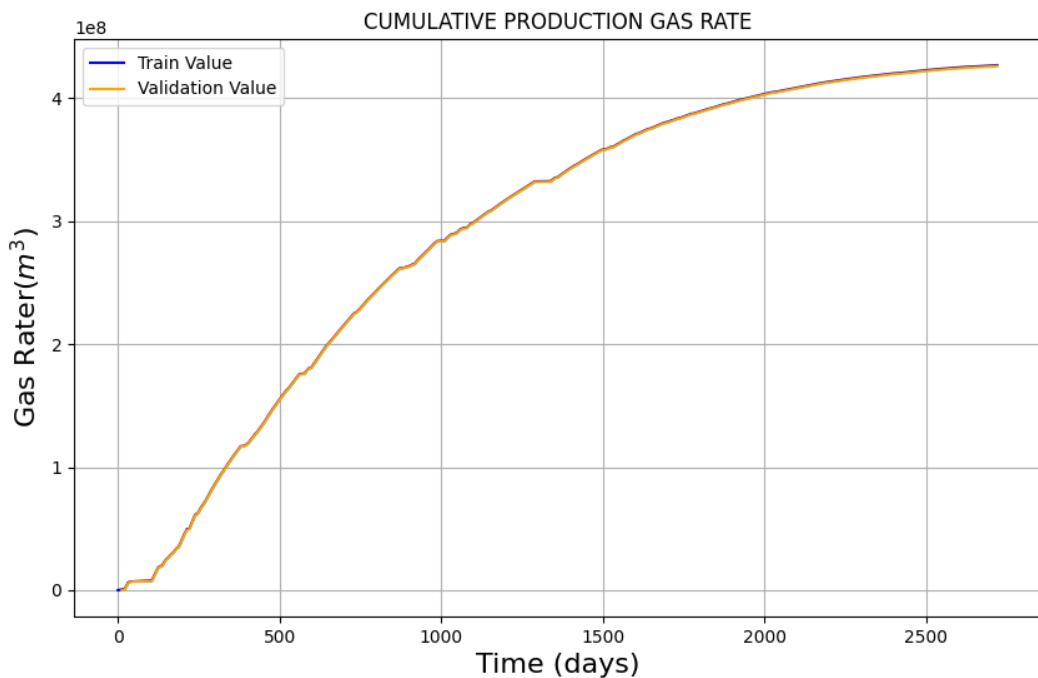
Hình 3-36: Sản lượng dầu khai thác tích lũy theo thời gian – LSTM



Hình 3-37: Dự báo sản lượng khí khai thác theo mô hình LSTM



Hình 3-38: Sản lượng khí khai thác dự báo so với giá trị thực tế – LSTM



Hình 3-39: Sản lượng khí khai thác tích lũy theo thời gian – LSTM

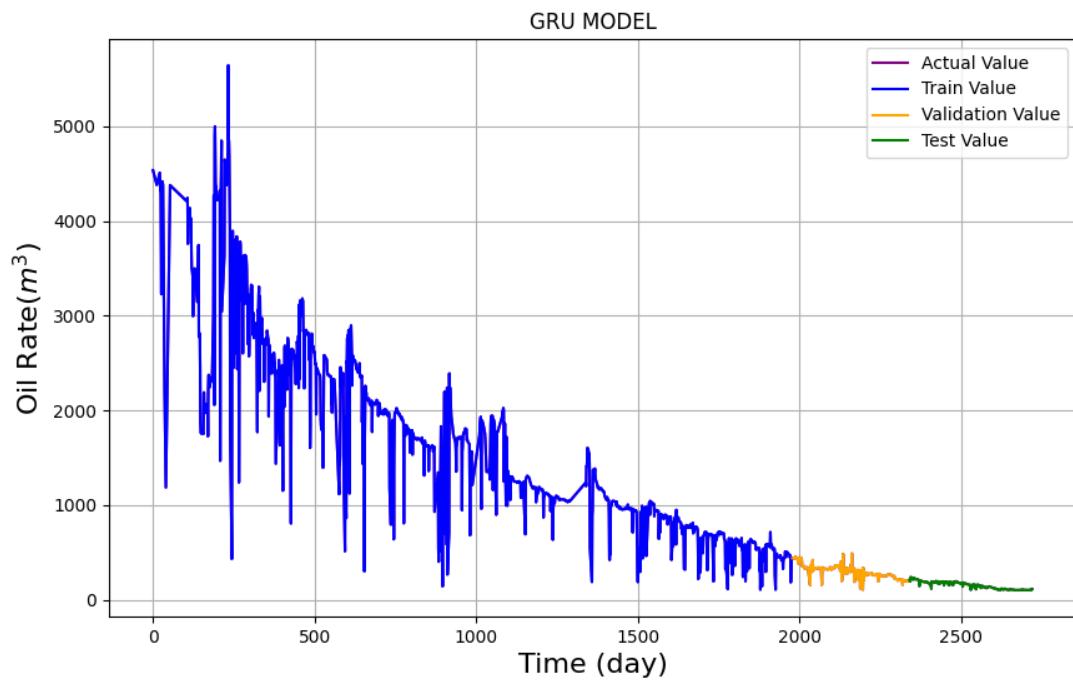
3.7.2.3 Đánh giá mô hình nút hồi quy có cổng (Gated Recurrent Unit)

Bảng 3-12: Bảng đánh giá mô hình nút hồi quy có cổng dự báo sản lượng dầu khai thác

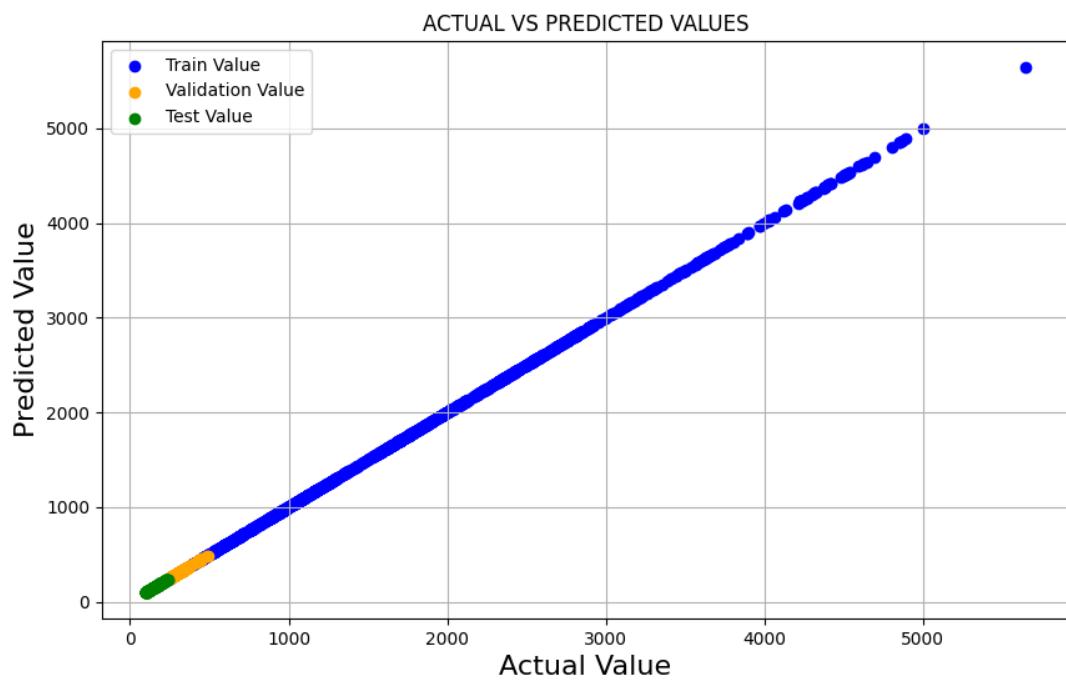
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	3.6040	3.5498	1.0000
Xác thực	3.6564	3.5375	0.9971
Kiểm tra	3.6564	3.5375	0.9909

Bảng 3-13: Bảng đánh giá mô hình nút hồi quy có cổng dự báo sản lượng khí khai thác

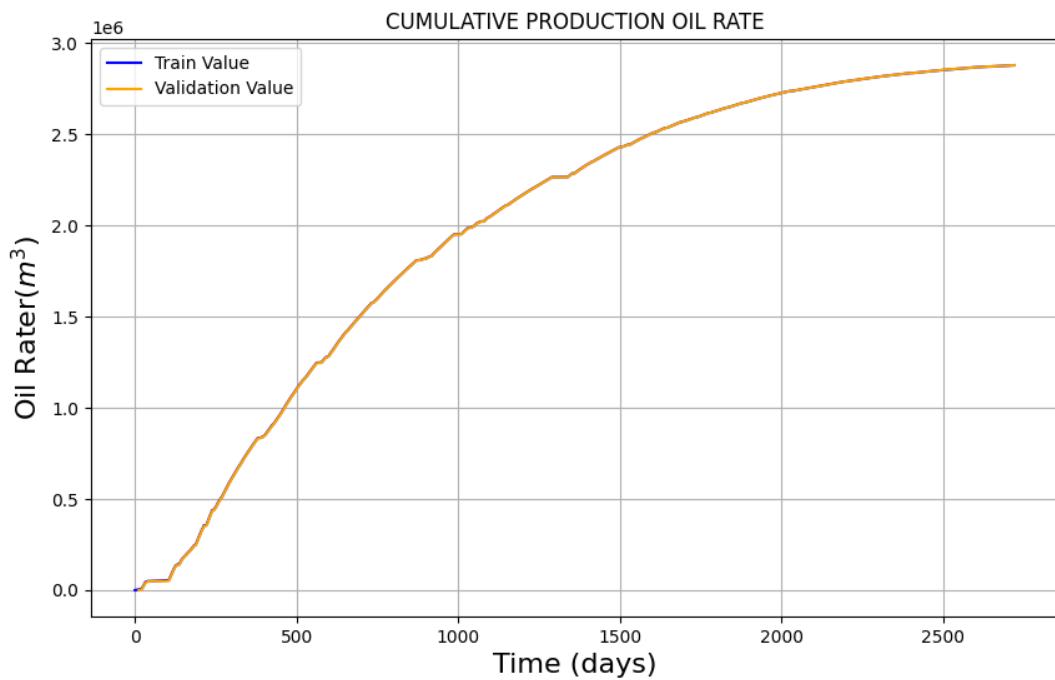
Tập dữ liệu	RMSE	MAE	R ²
Huấn luyện	188.9061	186.6283	1.0000
Xác thực	183.0804	177.3765	0.9997
Kiểm tra	183.1262	177.4189	0.9988



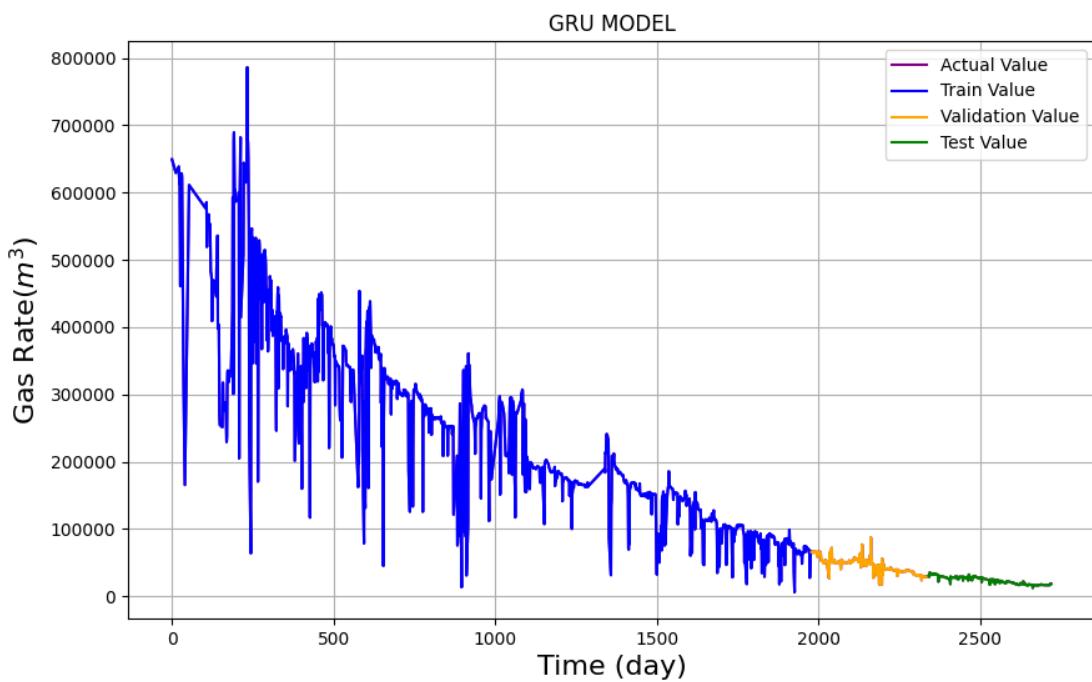
Hình 3-40: Dự báo sản lượng dầu khai thác theo mô hình GRU



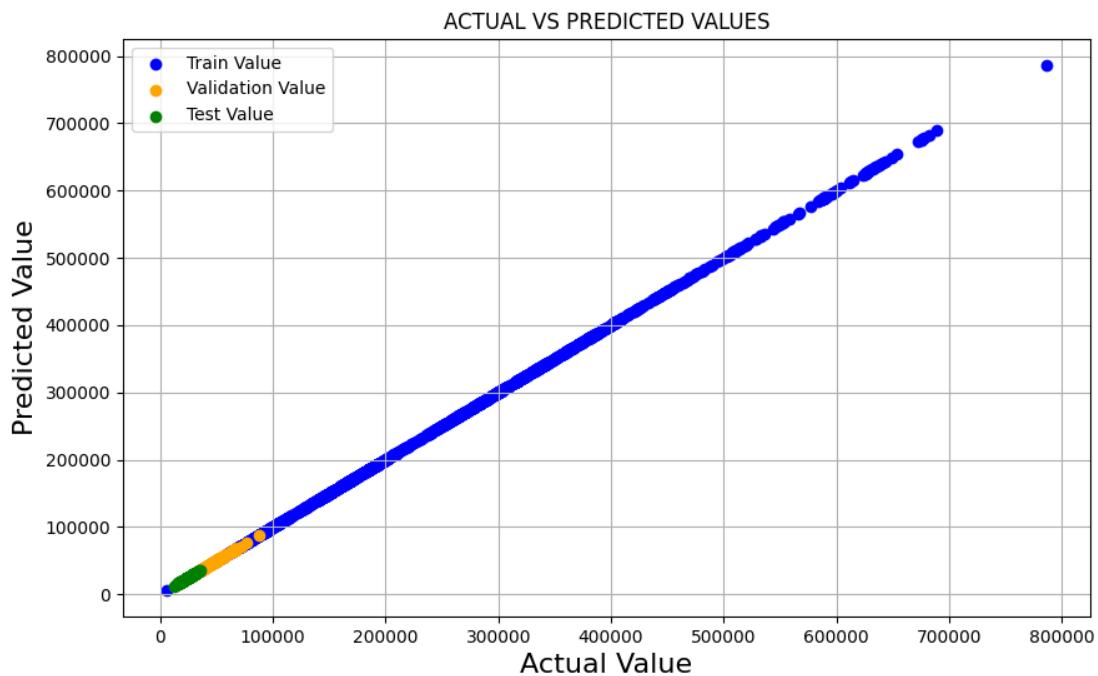
Hình 3-41: Sản lượng dầu khai thác dự báo so với giá trị thực tế – LSTM



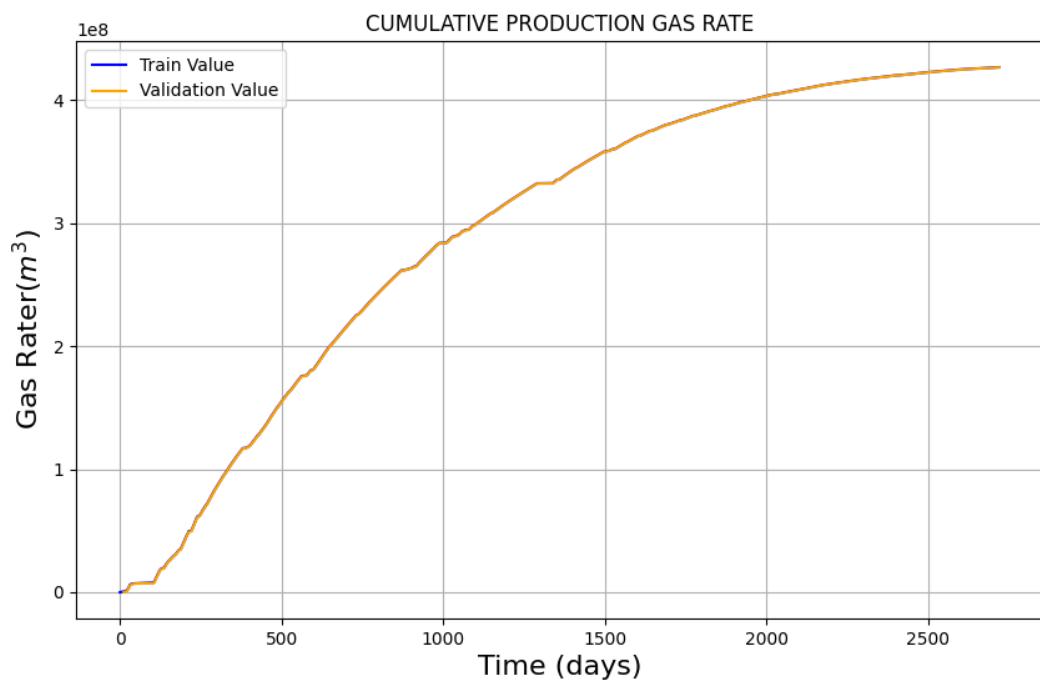
Hình 3-42: Sản lượng dầu khai thác tích lũy theo thời gian – GRU



Hình 3-43: Dự báo sản lượng khí khai thác theo mô hình GRU



Hình 3-44: Sản lượng khí khai thác dự báo so với giá trị thực tế – GRU



Hình 3-45: Sản lượng khí khai thác tích lũy theo thời gian – GRU

3.7.3 So sánh các mô hình học máy với học sâu, học máy và học sâu với đường cong suy giảm DCA và mạng nơ – ron nhân tạo ANN

Các mô hình học máy và học sâu cho kết quả gần như tương đương nhau, tuy nhiên chúng vẫn có một số khác biệt nhất định. Các bảng sau tổng hợp các kết quả đạt được sau quá trình phân tích, giá trị so sánh chính là các thông số đánh giá của pha kiểm tra.

Bảng 3-14: Bảng đánh giá tổng hợp các mô hình học máy và học sâu trong dự báo sản lượng dầu khai thác

MÔ HÌNH	Huấn luyện			Xác thực			Kiểm tra		
	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²
LN	372.7573	14.6724	0.9157	366.9816	14.3109	0.9266	12.0054	2.9951	0.9070
SVR	375.0697	14.0125	0.9129	373.4134	13.8088	0.9152	15.8188	3.5612	0.8444
SRNN	2.1230	1.7535	0.9999	2.9087	2.1874	0.9982	2.9087	2.1874	0.9942
LSTM	5.3721	2.1311	0.9999	0.6896	0.4330	0.9999	0.3170	0.2882	0.9999
GRU	3.6040	3.5498	0.9999	3.6564	3.5375	0.9971	3.6564	3.5375	0.9909

Bảng 3-15: Bảng đánh giá tổng hợp các mô hình học máy và học sâu trong dự báo sản lượng khí khai thác

MÔ HÌMINH	Huấn luyện			Xác thực			Kiểm tra		
	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²
LN	52316.4032	174.8368	0.9132	55749.0699	174.8370	0.9189	1847.1564	38.1661	0.8750
SVR	53080.5774	170.2241	0.9124	57069.8812	178.5682	0.8870	2397.7061	42.7614	0.8132
SRNN	203.6990	202.2978	0.9999	206.3152	202.6351	0.9997	206.3152	202.6351	0.9985
LSTM	111.0014	76.4531	0.9999	59.4249	58..4101	0.9999	58.5009	58.4232	0.9999
GRU	188.9061	186.6283	0.9999	183.0804	177.3765	0.9997	183.1262	177.4189	0.9988

Như thể hiện ở hai bảng trên, ta nhận thấy rằng trong cả 2 trường hợp ứng dụng học máy và học sâu để dự báo sản lượng khai thác dầu và khí, các thuật toán học sâu có nhiều ưu thế vượt trội hơn so với các thuật toán học máy. Do đó, ta chỉ cần chọn thuật toán tốt nhất để so sánh với các phương pháp dự báo khai thác khác, ở đây ta chọn LSTM.

Dưới đây là tổng hợp các đánh giá về mức độ hiệu quả của các mô hình. Phần trăm sai số được tính theo công thức:

$$\delta = \frac{\Delta x}{x} = \frac{|x_0 - x|}{x} \quad (3-6)$$

Bảng 3-16: So sánh mô hình LSTM với phương pháp DCA và ANN trong dự báo sản lượng dầu khai thác

Thông số đánh giá	Ký hiệu	LSTM	DCA	ANN
Sai số toàn phương trung bình khai căn	RMSE	4.5031	421.2365	165.0261
Sai số tuyệt đối trung bình	MAE	1.5994	218.3462	81.5997
Hệ số xác định	R ²	0.9999	0.9345	0.9876
Tổng sản lượng khai thác thực tế	Np		2,878,294	
Tổng sản lượng khai thác dự báo	Np'	2,878,453	3,011,331	2,873,882
Sai số tổng sản lượng khai thác	δ	$9.5169 \times 10^{-3}\%$	4.622%	0.1535%

Bảng 3-17: So sánh mô hình LSTM với phương pháp DCA và ANN trong dự báo sản lượng khai thác

Thông số đánh giá	Ký hiệu	LSTM	DCA	ANN
Sai số toàn phuong trung bình khai cǎn	RMSE	98.3106	57071,068	25672,6881
Sai số tuyệt đối trung bình	MAE	71.0361	33220,225	13099,8853
Hệ số xác định	R ²	0.9999	0,9300	0.9852
Tổng sản lượng khai thác thực tế	Np		426,599,669	
Tổng sản lượng khai thác dự báo	Np'	426,436,977	428,660,092	423,769,813
Sai số tổng sản lượng khai thác	δ	0.0338%	0.4830%	0.6678%

Các kết quả cho thấy, thuật toán học sâu mạng nơ – ron hồi quy ưu việt hơn hẳn phương pháp phân tích đường cong suy giảm truyền thống và mạng nơ ron – nhân tạo truyền thống (ANN). Nguyên nhân có thể kể đến là do bộ dữ liệu của ta có sự biến động mạnh tại nhiều thời điểm, tuy nhiên càng gần đến thời điểm đóng giếng, các số liệu khai thác đã dần ổn định hơn, trong giai đoạn này DCA vẫn có thể dự báo có độ chính xác khá cao so với thực tế, tuy nhiên vẫn không thể đạt được hiệu quả như LSTM. Một nguyên nhân khác có thể dẫn đến tình trạng không thể sử dụng DCA để dự báo khai thác là khi dữ liệu khai thác không tuân theo quy luật nào, cả ba mô hình Exponential, Harmonic và Hyperbolic đều không thể áp dụng được, khi đó phương pháp học máy và học sâu là sự lựa chọn hợp lý.

KẾT LUẬN VÀ KIẾN NGHỊ

1. Kết luận

Học máy, học sâu và trí tuệ nhân tạo ra đời rất sớm. Tuy nhiên, khi nó được đề xuất, dữ liệu dành cho công tác huấn luyện không nhiều; cho đến những năm gần đây với công nghệ dữ liệu Big Data, trí tuệ nhân tạo được phát triển mạnh mẽ. Không dừng ngoài xu hướng phát triển và cũng để khắc phục những nhược điểm có hữu của những phương dự báo truyền thống, ngành dầu khí cũng có nhiều nghiên cứu về ứng dụng trí tuệ nhân tạo trong dự báo khai thác dầu khí.

Dự báo lưu lượng khai thác dầu là một công việc cần thiết trong quá trình khai thác. Nhờ nó mà ta có thể dự đoán tương đối chính xác thời điểm thích hợp để đóng giếng hoặc tiến hành các biện pháp thu hồi dầu tăng cường nhằm tối đa hóa sản lượng khai thác và tối thiểu hóa chi phí vận hành. Luận văn có sự kết hợp nhiều phương pháp dự báo khai thác dầu khí sử dụng các thuật toán học máy và học sâu, sau khi phân tích và đưa ra kết quả, có thể kết luận như sau:

- Các thuật toán học máy và học sâu được sử dụng đều dựa vào các quan hệ đại số trên tập dữ liệu gốc. Việc tìm ra những mối quan hệ có trong bộ số liệu cũng chính là tìm ra những tác động vật lý, hóa học của via ảnh hưởng đến lưu lượng khai thác.
- Kết quả đầu ra của một thuật toán học máy, học sâu phụ thuộc nhiều vào tập dữ liệu đầu vào. Nếu tập đầu vào lộn xộn, tỉ lệ phân mảnh lớn hay thiếu đồng nhất thì dự báo khai thác có thể không đáng tin cậy.
- Các thuật toán học máy, học sâu chỉ có thể được dự báo trong ngắn hạn. Khi thời gian dự báo mà ta yêu cầu quá dài, ta sẽ không đánh giá được các tác động ngoại sinh ảnh hưởng lên mô hình, từ đó mô hình sẽ có nhiều sai lệch.
- Đối với các mô hình hồi quy tuyến tính và hồi quy véc – tơ hỗ trợ, mặc dù hệ số xác định có thể chấp nhận được, nhưng sai số là quá lớn nếu so với các mô hình mạng nơ – ron hồi quy. Điều này chứng tỏ các mô hình này không phù hợp với những bộ số liệu có sự biến động mạnh như dữ liệu khai thác giếng F14 của mỏ Volve.

- Phương pháp học sâu mạng nơ – ron hồi quy vượt trội hơn so với phương pháp dự báo truyền thống và mạng nơ – ron nhân tạo. Dù bộ dữ liệu có biên độ biến động khá lớn, nó vẫn đáp ứng tốt các tiêu chí: sai số thấp, hệ số xác định cao và là mô hình đơn giản (không phụ thuộc vào các đặc trưng khác để dự báo).

2. Kiến nghị

Các phương pháp học máy, học sâu chỉ đưa ra các dự báo dựa trên số liệu lịch sử khai thác. Do đó, nó không có bất kỳ phản ánh nào về trữ lượng tiềm năng khai thác còn lại trong vỉa khi ta chuyển sang một chế độ khai thác mới. Ta cần phải thực hiện các công tác khảo sát địa chất, địa vật lý và mô hình hóa vỉa để đánh giá đúng đắn tiềm năng của vỉa và tiến hành chuyển đổi sang một chế độ khai thác mới nhằm thu hồi dầu tăng cường, nếu khả thi.

Việc phân tích dữ liệu khai thác cần được tiến hành thường xuyên, kết hợp nhiều phương pháp để tìm ra ưu và nhược điểm của mỗi phương pháp tương ứng. Việc ứng dụng trí tuệ nhân tạo sẽ tối ưu hóa được chi phí vận hành khai thác, vì thế cần tiếp tục áp dụng công nghệ này vào hoạt động khai thác cũng như các hoạt động khác của ngành dầu khí.

TÀI LIỆU THAM KHẢO

- [1] R. G. Agatwal, D. C. Gardner, and S. W. Kleinsteiber, “Analyzing Well production data using combined type curve and decline curve analysis concepts,” 1996.
- [2] R. J. Boomer and T. Exploration, “SPE 30202 Predicting Production Using a Neural Network (Artificial Intelligence Beats Human Intelligence) Society of Petroleum Engineers,” pp. 195–204, 1995.
- [3] J. Sun, X. Ma, and M. Kazi, “Comparison of Decline Curve Analysis DCA with Recursive Neural Networks RNN for Production Forecast of Multiple Wells,” *SPE West. Reg. Meet. Proc.*, vol. 2018-April, 2018, doi: 10.2118/190104-ms.
- [4] Y. Li, R. Sun, and R. Horne, “Deep learning for well data history analysis,” *Proc. - SPE Annu. Tech. Conf. Exhib.*, vol. 2019-Sept, no. October, 2019, doi: 10.2118/196011-ms.
- [5] H. Alimohammadi, H. Rahmanifard, and N. Chen, “Multivariate time series modelling approach for production forecasting in unconventional resources,” *Proc. - SPE Annu. Tech. Conf. Exhib.*, vol. 2020-Octob, pp. 1–13, 2020, doi: 10.2118/201571-ms.
- [6] H. Sun, *Advanced Production Decline Analysis and Application*. 2015.
- [7] T. H. Vu, “Machine Learning Cơ bản,” 2020.
- [8] D. Jap, M. Stöttinger, and S. Bhasin, “Support vector regression,” no. November 2007, pp. 1–8, 2015, doi: 10.1145/2768566.2768568.
- [9] R. M. Schmidt, “Recurrent Neural Networks (RNNs): A gentle Introduction and Overview,” no. 1, pp. 1–16, 2019, [Online]. Available: <http://arxiv.org/abs/1912.05911>.
- [10] R. Dey and F. M. Salem, “Gate-Variants of Gated Recurrent Unit (GRU),” vol. 784.

- [11] S. Sharma, S. Sharma, and A. Anidhya, “Understanding Activation Functions in Neural Networks,” *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 12, pp. 310–316, 2020.
- [12] S. Ruder, “An overview of gradient descent optimization algorithms,” pp. 1–14, 2016.

PHỤ LỤC A

DỰ BÁO KHAI THÁC DẦU KHÍ MỎ VOLVE BẰNG PHƯƠNG PHÁP PHÂN TÍCH ĐƯỜNG CONG SUY GIẢM TRÊN PHẦN MỀM EXCEL

1. Quy trình phân tích đường cong suy giảm với mô hình Exponential

Bước 1: Vẽ đồ thị $\log(q)$ với t trên thang đo Decartes và vẽ đường thẳng hồi quy tuyến tính của tập dữ liệu trên.

Bước 2: Đường thẳng hồi quy cắt trục tung tại điểm $t_0 = 0$. Từ đó, ta dễ dàng tính được giá trị lưu lượng ban đầu q_i tại t_0 bằng phương trình đường thẳng hồi quy ở bước 1.

Bước 3: Chọn một điểm bất kì (t_i, q_i) trên đường hồi quy, ta dễ dàng tính được tốc độ suy giảm D_i với công thức:

$$D_i = \frac{1}{t} \ln\left(\frac{q_i}{q_t}\right) \quad (\text{CTPL 1})$$

Bước 4: Tính toán lưu lượng khai thác theo thời gian và lưu lượng khai thác tích lũy với công thức:

$$q = q_i e^{-D_i t} \quad (\text{CTPL 2})$$

$$Q = \frac{q_i - q_{ab}}{D_i} \quad (\text{CTPL 3})$$

Bước 5: Đánh giá các kết quả đạt được.

2. Quy trình phân tích đường cong suy giảm với mô hình Harmonic

Bước 1: Vẽ đồ thị $1/q$ với t trên thang đo Decartes và vẽ đường thẳng hồi quy tuyến tính của tập dữ liệu trên.

Bước 2: Đường thẳng hồi quy cắt trục tung tại điểm $t_0 = 0$. Từ đó, ta dễ dàng tính được giá trị lưu lượng ban đầu q_i tại t_0 bằng phương trình đường thẳng hồi quy ở bước 1.

Bước 3: Chọn một điểm bất kỳ $(t_i, 1/q_i)$ trên đường hồi quy, ta dễ dàng tính được tốc độ suy giảm D_i với công thức:

$$D_i = \frac{\frac{q_1}{q_i} - 1}{\frac{t_2 - t_1}{t_2 - t_i}} \quad (\text{CTPL 4})$$

Bước 4: Tính toán lưu lượng khai thác theo thời gian và lưu lượng khai thác tích lũy với công thức:

$$q = \frac{q_i}{1 + D_i t} \quad (\text{CTPL 5})$$

$$Q = \frac{q_i}{D_i} \ln\left(\frac{q_i}{q_t}\right) \quad (\text{CTPL 6})$$

Bước 5: Đánh giá các kết quả đạt được.

3. Quy trình phân tích đường cong suy giảm với mô hình Hyperbolic

Bước 1: Vẽ đồ thị q với t trên thang Decartes và vẽ đường hồi quy của tập dữ liệu trên.

Bước 2: Đường thẳng hồi quy cắt trục tung tại điểm $t_0 = 0$. Từ đó, ta dễ dàng tính được giá trị lưu lượng ban đầu q_i tại t_0 bằng phương trình đường thẳng hồi quy ở bước 1.

Bước 3: Chọn một điểm ở khoảng cuối đường hồi quy, ghi lại tọa độ điểm (t_2, q_2)

Bước 4: Xác định điểm ở giữa đường hồi quy (t_1, q_1) , trong đó q_1 được xác định theo công thức:

$$q_1 = \sqrt{q_i q_2} \quad (\text{CTPL 7})$$

Bước 5: Giải phương trình xác định b bằng phương pháp lặp tiếp tuyến theo công thức:

$$f(b) = t_2 \left(\frac{q_i}{q_1} \right)^b - t_1 \left(\frac{q_i}{q_2} \right)^b - (t_2 - t_1) = 0 \quad (\text{CTPL 8})$$

Bước 6: Xác định D_i theo công thức:

$$D_i = \frac{\left(\frac{q_i}{q_2} \right)^b - 1}{bt_2} \quad (\text{CTPL 9})$$

Bước 7: Tính toán lưu lượng khai thác theo thời gian và lưu lượng khai thác tích lũy với công thức:

$$q = \frac{q_i}{(1 + bD_i t)^{\frac{1}{b}}} \quad (\text{CTPL 10})$$

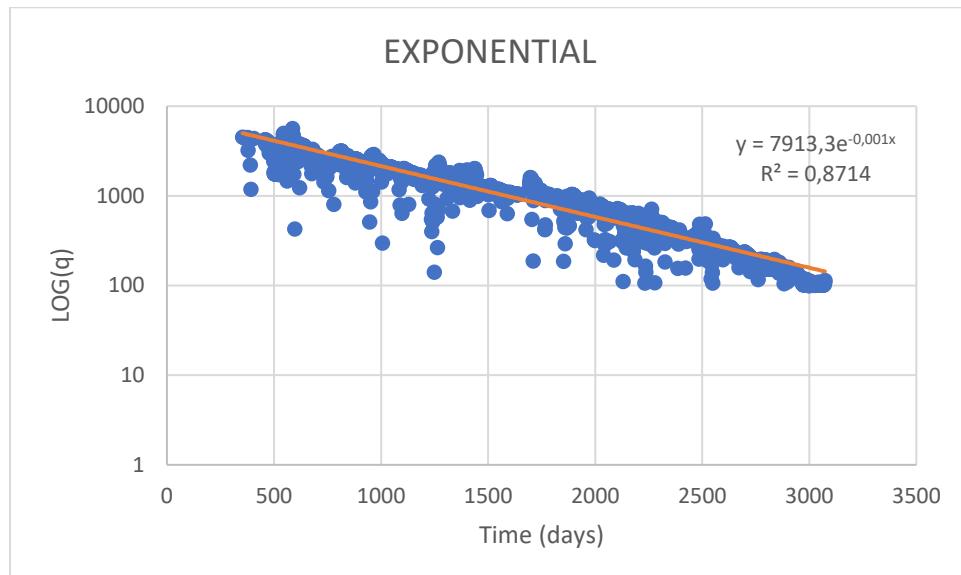
$$Q = \frac{q_i}{(1-b)D_i} \left[1 - \left(\frac{q_t}{q_i} \right)^{1-b} \right] \quad (\text{CTPL 11})$$

Bước 8: Đánh giá các kết quả đạt được.

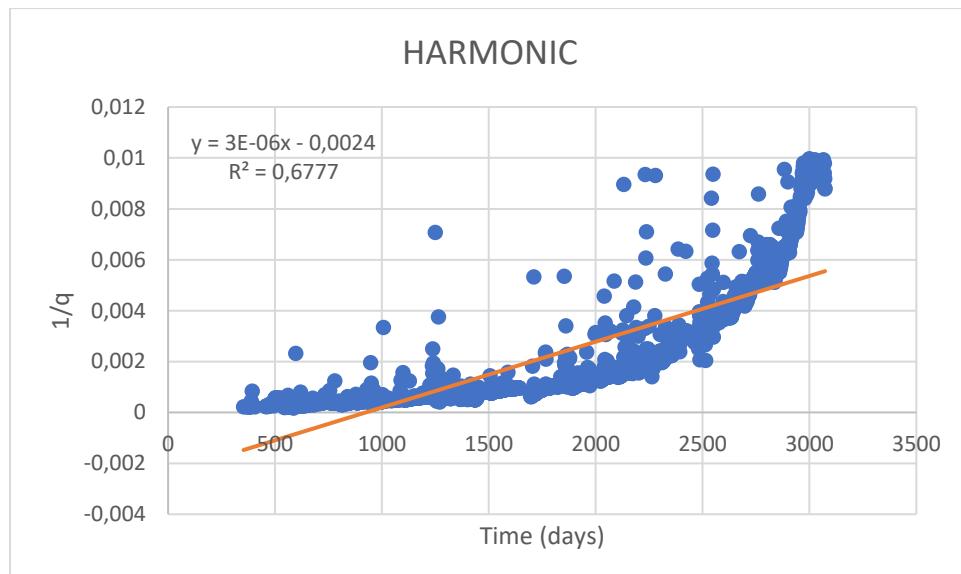
4. Dự báo lưu lượng dầu khai thác bằng phương pháp DCA

4.1 Dự báo lưu lượng dầu khai thác bằng phương pháp DCA

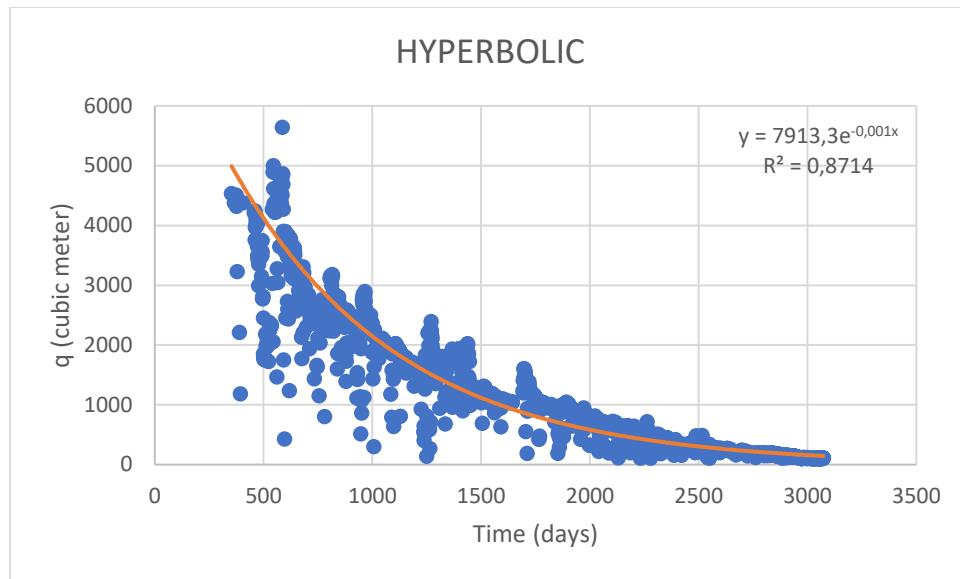
Bước 1: Vẽ các dạng đồ thị Exponential, Harmonic, Hyperbolic.



Hình PL 1: Đồ thị semi – log của log(q) và t trong mô hình Exponential



Hình PL 2: Đồ thị 1/q với t trong mô hình Harmonic



Hình PL 3: Đồ thị q_t với t trong mô hình Hyperbolic

Nhận xét: Trong 3 mô hình, dễ dàng thấy được mô hình Harmonic không phù hợp với bộ dữ liệu, do đó ta có thể loại mô hình này khỏi các tính toán tiếp theo.

Bước 2: Xác định lưu lượng ban đầu q_i tại thời điểm $t=0$ dựa vào phương trình hồi quy:

$$q_i = 7913.3e^{-0.001x} = 7913.3e^{-0.001 \times 0} = 7913.3(m^3)$$

Bước 3: Chọn giá trị nằm ở gần cuối đường hồi quy $(t_2, q_2) = (1712, 891.56)$.

Bước 4: Tính D_i

- Đối với mô hình Exponential:

Ta tiến hành tính tốc độ suy giảm D_i với công thức:

$$D_i = \frac{1}{t_2} \ln \frac{q_i}{q_2} = \frac{1}{1712} \ln \frac{7913.3}{891.56} = 0.0012753$$

- Đối với mô hình Hyperbolic:

Ta tìm điểm (t_1, q_1) bằng công thức:

$$q_1 = \sqrt{q_i q_2} = \sqrt{7913.3 \times 891.56} = 2656.159(m^3)$$

$$q_1 = 7913.3 e^{-0.001 \times t_1}$$

$$\Rightarrow t_1 = \frac{1}{-0.001} \ln \frac{q_1}{7913.3} = \frac{1}{-0.001} \ln \frac{2656.159}{7913.3} = 1091.664(m^3)$$

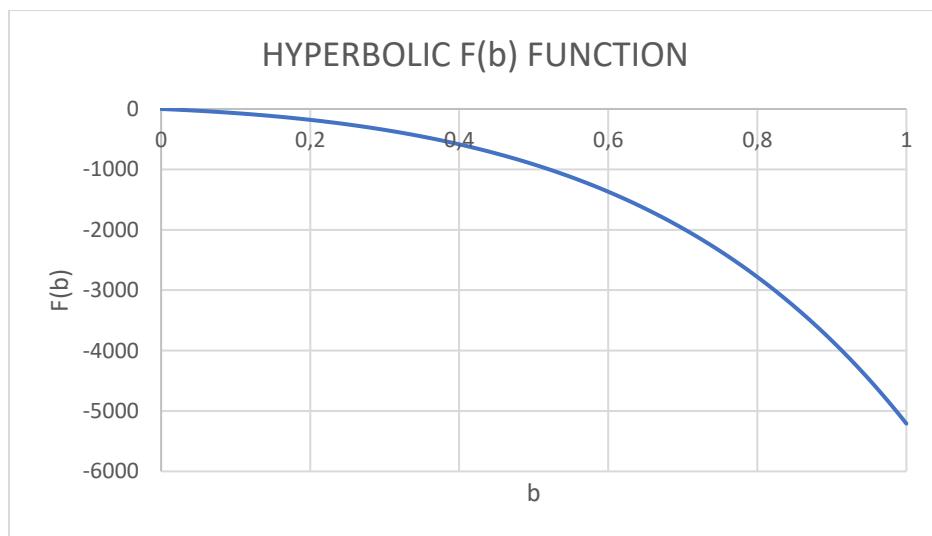
Để tìm được D_i đối với phương pháp Hyperbolic, ta cần phải tính hệ số b, cho bởi công thức:

$$f(b) = t_2 \left(\frac{q_i}{q_1} \right)^b - t_1 \left(\frac{q_i}{q_2} \right)^b - (t_2 - t_1) = 0$$

$$\Leftrightarrow 1712 \left(\frac{7913.3}{2656.159} \right)^b - 1091.664 \left(\frac{7913.3}{891.56} \right)^b - (1712 - 1091.664) = 0$$

$$\Leftrightarrow 1712 \times 2.9792^b - 1091.664 \times 8.8758^b - 620.336 = 0$$

Tuy nhiên phương trình trên vô nghiệm trên khoảng $(0, 1)$, nên ta không tìm được b và do đó cũng không thể tìm được D_i tương ứng. Các bước tiếp theo chỉ áp dụng riêng cho mô hình Exponential và ta chấp nhận Exponential là mô hình duy nhất có thể dùng để dự báo khai thác dầu khí trong phương pháp DCA với bộ dữ liệu mà ta có được.



Hình PL 4: Phương trình $f(b)$ vô nghiệm trên khoảng $(0, 1)$

Bước 5: Tính lưu lượng dầu khai thác theo công thức: $q = q_i e^{-D_i t}$

Kết quả tính toán được thể hiện trong bảng sau:

Bảng PL 1: Kết quả dự báo khai thác dầu mỏ Volve với phương pháp DCA – mô hình Exponential

STEP	OIL	Cumulative	EXPONENTIAL			
			q	N	SQR ERROR	ABS ERROR
353	4535,43	4535,43	5044,816	5044,816	259473,87	509,38577
365	4379,88	8915,31	4968,199	10013,01	346119,48	588,3192
374	4509,07	13424,38	4911,501	14924,52	161950,98	402,43134
376	4319,02	17743,4	4898,99	19823,51	336365,14	579,96995
377	4417,66	22161,06	4892,746	24716,25	225706,91	475,08621
379	3226,61	25387,67	4880,283	29596,54	2734633	1653,6726
380	4411,9	29799,57	4874,063	34470,6	213594,36	462,1627
...
3068	106,19	2877749,39	158,1676	3010544	2701,6668	51,977561
3069	106,3	2877855,69	157,966	3010701	2669,3732	51,665977
3070	102,09	2877957,78	157,7647	3010859	3099,6667	55,67465
3071	113,38	2878071,16	157,5636	3011017	1952,1887	44,18358
3072	108,84	2878180	157,3628	3011174	2354,4588	48,522766
3073	113,84	2878293,84	157,1622	3011331	1876,8137	43,322208

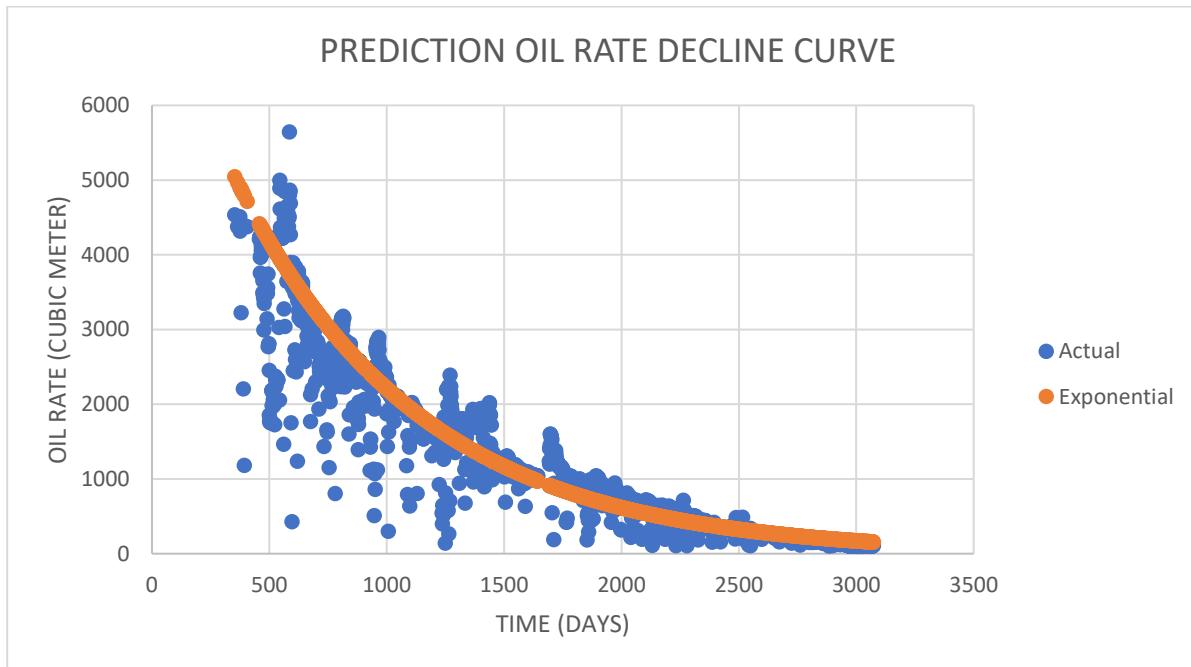
Trong đó, STEP là mốc thời gian, OIL là lưu lượng khai thác của dầu, Cumulative là sản lượng khai thác thực tế tích lũy, q là lưu lượng khai thác dự đoán của dầu, N là sản lượng khai thác tích lũy dự báo tính theo mốc thời gian của dầu, SQR ERROR là sai số bình phương, ABS ERROR là sai số tuyệt đối.

Bước 6: Đánh giá các kết quả đạt được

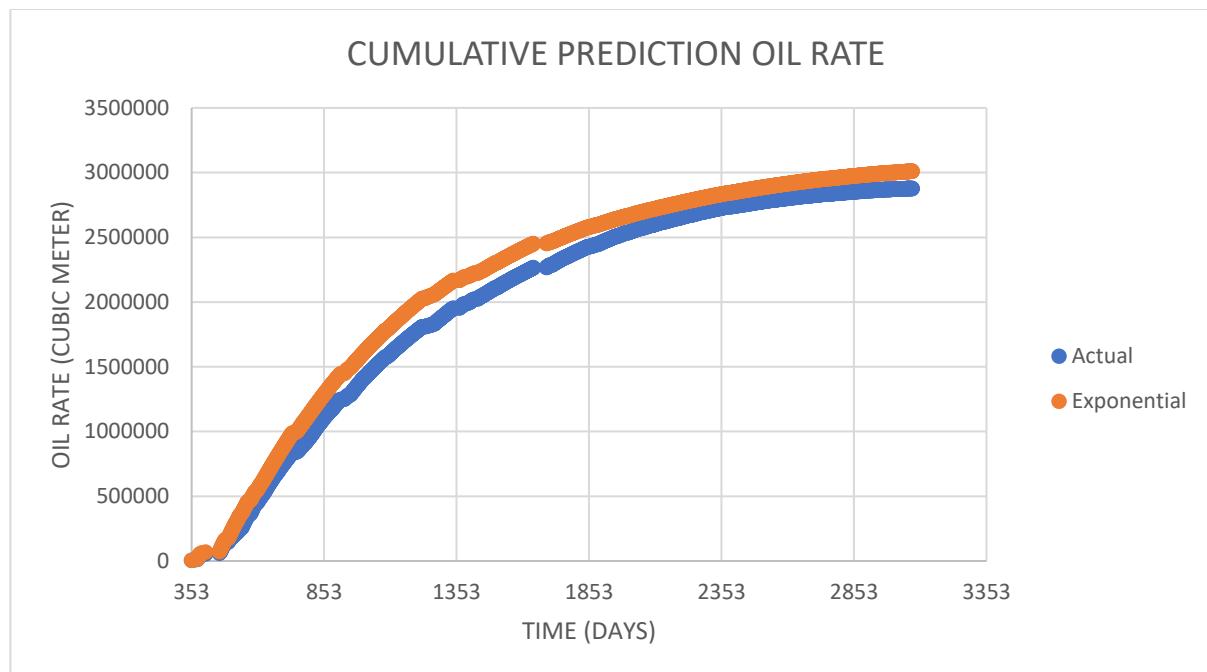
Dựa vào kết quả tính toán trên Excel, ta có được bảng sau:

Bảng PL 2: Sai số đánh giá mô hình Exponential – phương pháp DCA trong dự báo khai thác dầu mỏ Volve

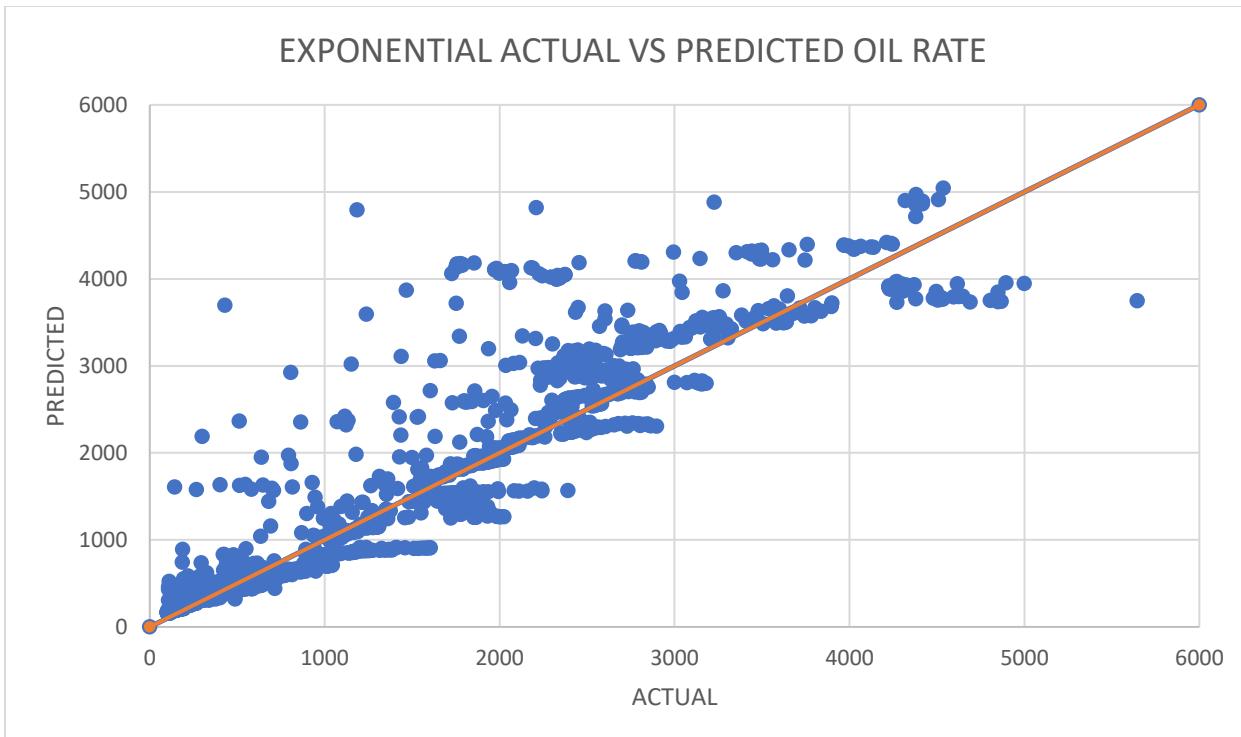
Sai số toàn phương trung bình	RMSE	421,23648
Sai số tuyệt đối trung bình	MAE	218,34624
Hệ số xác định	R ²	0,9344844



Hình PL 5: Đồ thị dự báo lưu lượng khai thác dầu theo thời gian trong phương pháp Exponential – DCA



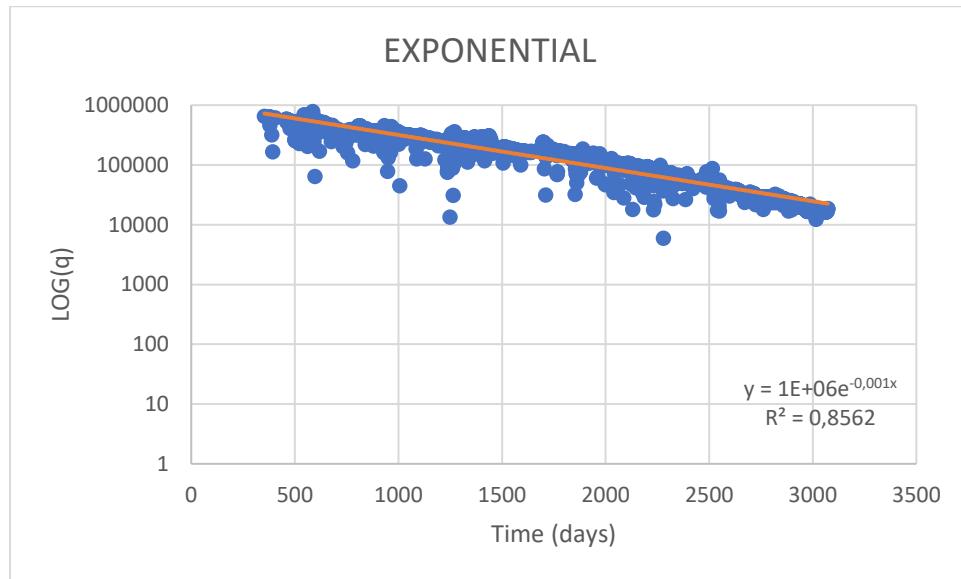
Hình PL 6: Đồ thị dự báo lưu lượng khai thác dầu tích lũy theo thời gian trong phương pháp Exponential – DCA



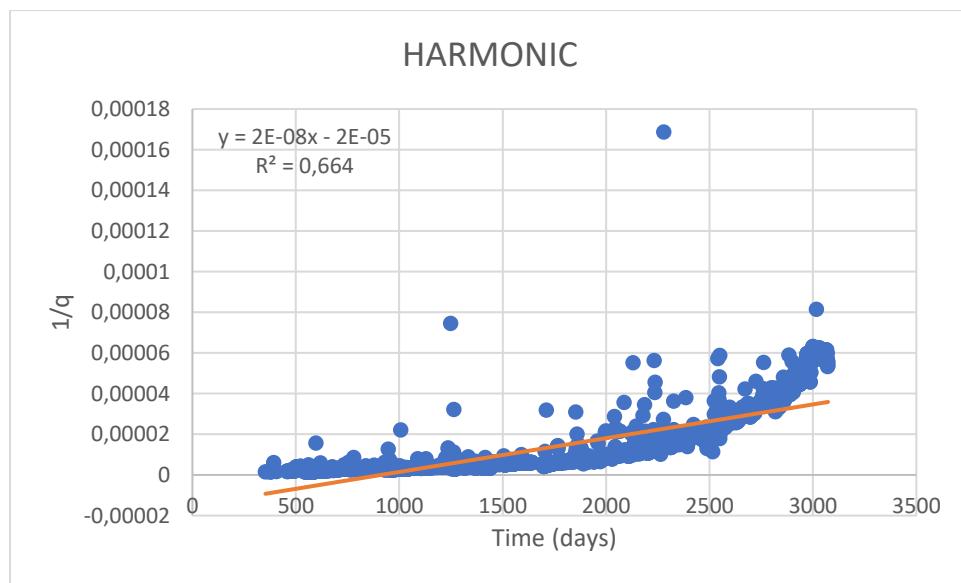
Hình PL 7: Giá trị khai thác dầu dự báo so với giá trị khai thác dầu thực tế trong phương pháp Exponential - DCA

4.2 Dự báo lưu lượng khí khai thác bằng phương pháp DCA

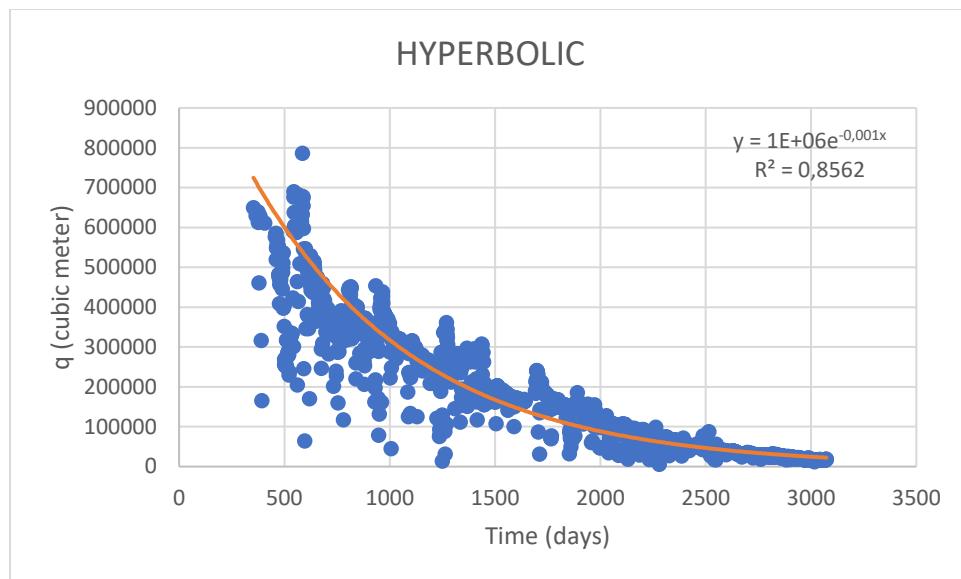
Bước 1: Vẽ các dạng đồ thị Exponential, Harmonic, Hyperbolic.



Hình PL 8: Đồ thị semi – log của $\log(q)$ và t trong mô hình Exponential



Hình PL 9: Đồ thị $1/q$ với t trong mô hình Harmonic



Hình PL 10: Đồ thị q_t với t trong mô hình Hyperbolic

Nhận xét: Trong 3 mô hình, dễ dàng thấy được mô hình Harmonic không phù hợp với bộ dữ liệu, do đó ta có thể loại mô hình này khỏi các tính toán tiếp theo.

Bước 2: Xác định lưu lượng ban đầu q_i tại thời điểm $t = 0$ dựa vào phương trình hồi quy:

$$q_i = 10^6 e^{-0.001x} = 10^6 e^{-0.001 \times 0} = 10^6 (m^3)$$

Bước 3: Chọn giá trị nằm ở gần cuối đường hồi quy $(t_2, q_2) = (1712, 134483.5)$.

Bước 4: Tính D_i

- Đối với mô hình Exponential:

Ta tiến hành tính tốc độ suy giảm D_i với công thức:

$$D_i = \frac{1}{t_2} \ln \frac{q_i}{q_2} = \frac{1}{1712} \ln \frac{10^6}{134483.5} = 0.0011719$$

- Đối với mô hình Hyperbolic:

Ta tìm điểm (t_1, q_1) bằng công thức:

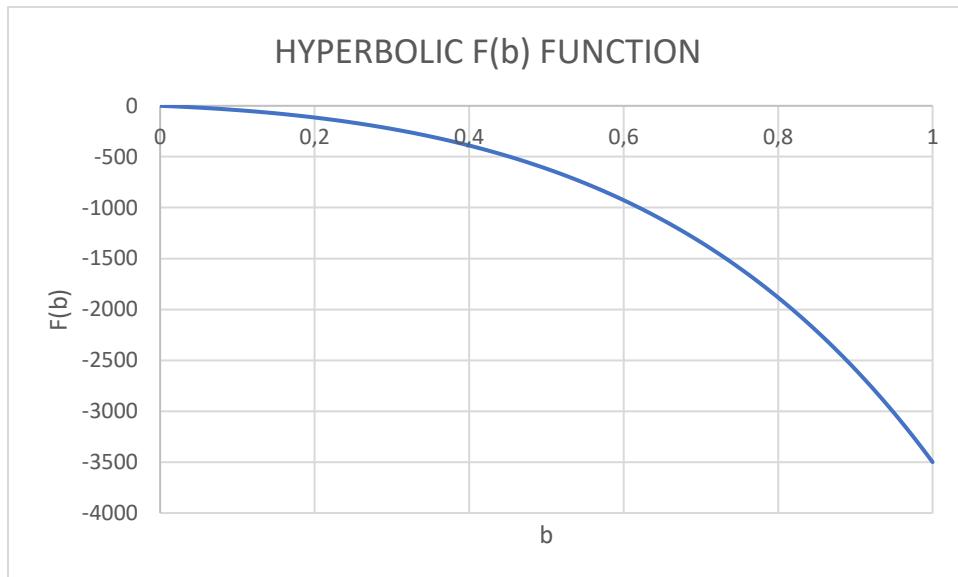
$$q_1 = \sqrt{q_i q_2} = \sqrt{10^6 \times 134483.5} = 366719.92(m^3)$$

$$\begin{aligned} q_1 &= 10^6 e^{-0.001 t_1} \\ \Rightarrow t_1 &= \frac{1}{-0.001} \ln \frac{q_1}{10^6} = \frac{1}{-0.001} \ln \frac{366719.92}{10^6} = 1003.1569(m^3) \end{aligned}$$

Để tìm được D_i đối với phương pháp Hyperbolic, ta cần phải tính hệ số b:

$$\begin{aligned} f(b) &= t_2 \left(\frac{q_i}{q_1} \right)^b - t_1 \left(\frac{q_i}{q_2} \right)^b - (t_2 - t_1) = 0 \\ \Leftrightarrow 1712 &\left(\frac{10^6}{366719.92} \right)^b - 1003.1569 \left(\frac{10^6}{134483.5} \right)^b - (1712 - 1003.1569) = 0 \\ \Leftrightarrow 1712 \times 2.7269^b &- 1003.1569 \times 7.4359^b - 708.8431 = 0 \end{aligned}$$

Tuy nhiên phương trình trên vô nghiệm trên khoảng $(0, 1)$, nên ta không tìm được b và do đó cũng không thể tìm được D_i tương ứng. Các bước tiếp theo chỉ áp dụng riêng cho mô hình Exponential và ta chấp nhận Exponential là mô hình duy nhất có thể dùng để dự báo khai thác dầu khí trong phương pháp DCA với bộ dữ liệu mà ta có được.



Hình PL 11: Phương trình $f(b)$ vô nghiệm trên khoảng $(0, 1)$

Bước 5: Tính lưu lượng dầu khai thác theo công thức: $q = q_i e^{-D_i t}$

Kết quả tính toán được thể hiện trong bảng sau:

Bảng PL 3: Kết quả dự báo khai thác khí mỏ Volve với phương pháp DCA – mô hình Exponential

STEP	GAS	Cumulative	EXPONENTIAL			
			q	N	SQR ERROR	ABS ERROR
353	649388,1	649388,07	661209,2	661209,2	139738746	11821,114
365	629307,3	1278695,41	651975,7	1313185	513855081	22668,372
374	638750,2	1917445,58	645135,3	1958320	40770190	6385,1539
376	612912,6	2530358,2	643625	2601945	943250932	30712,391
377	625514	3155872,21	642871,2	3244816	301271361	17357,17
379	460948	3616820,22	641366,2	3886183	3,255E+10	180418,16
380	628668,3	4245488,49	640615	4526798	142723949	11946,713
...
3068	17427,78	426510170	27448,97	4,29E+08	100424175	10021,186
3069	17541,2	426527711	27416,82	4,29E+08	97527819	9875,6174
3070	16681,29	426544392	27384,71	4,29E+08	114563116	10703,416
3071	18753,12	426563146	27352,63	4,29E+08	73951614	8599,5124
3072	17979,28	426581125	27320,6	4,29E+08	87260190	9341,3163
3073	18543,76	426599669	27288,6	4,29E+08	76472187	8744,8377

Các đại lượng của bảng tương tự như các đại lượng trong bảng 3.1, chỉ thay thế duy nhất một đại lượng là GAS cho OIL.

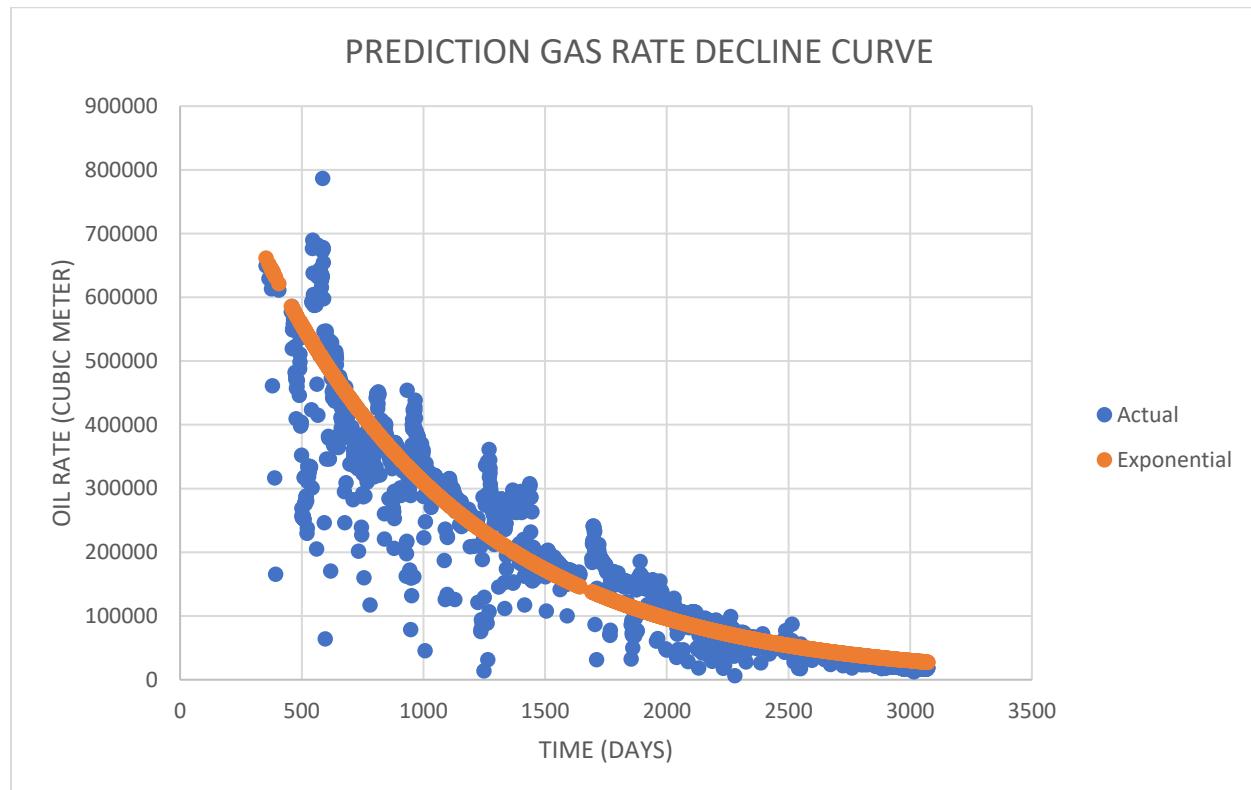
Bước 6: Đánh giá các kết quả đạt được

Để đánh giá tính hiệu quả của mô hình, ta hiện hành đánh giá các thông số: sai số toàn phương trung bình, sai số tuyệt đối trung bình và hệ số tương quan. Các công thức tính sai số tương tự như khi ta đánh giá hiệu quả dự báo khai thác dầu.

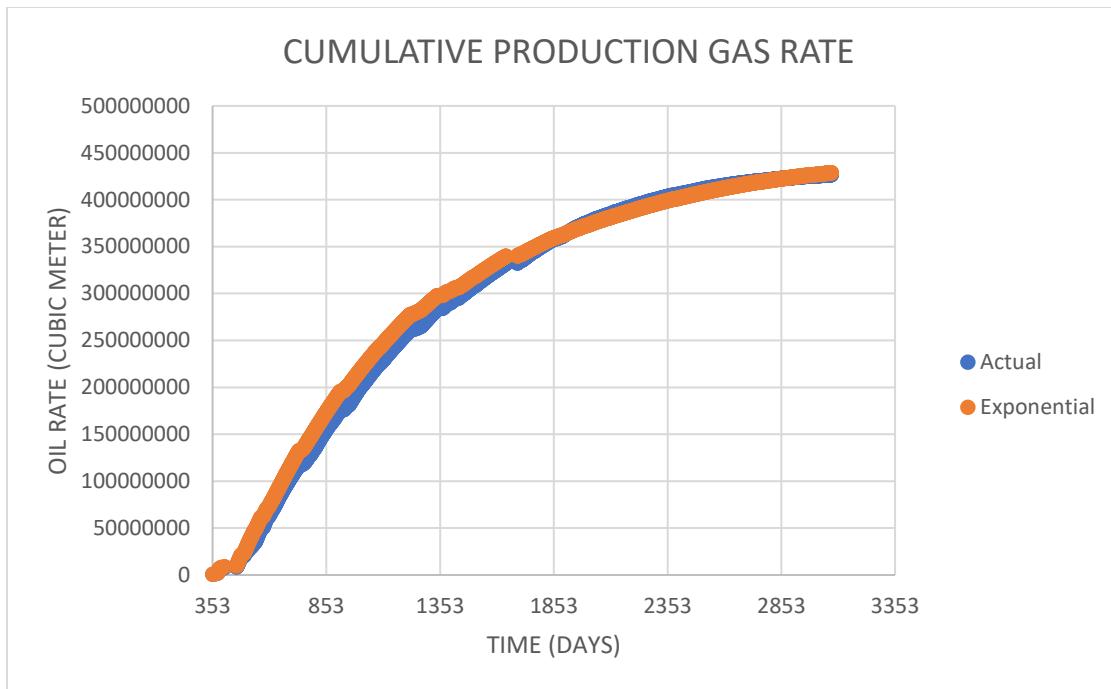
Dựa vào kết quả tính toán trên Excel, ta có được bảng sau:

Bảng PL 4: Sai số đánh giá mô hình Exponential – phương pháp DCA trong dự báo khai thác khí mỏ Volve

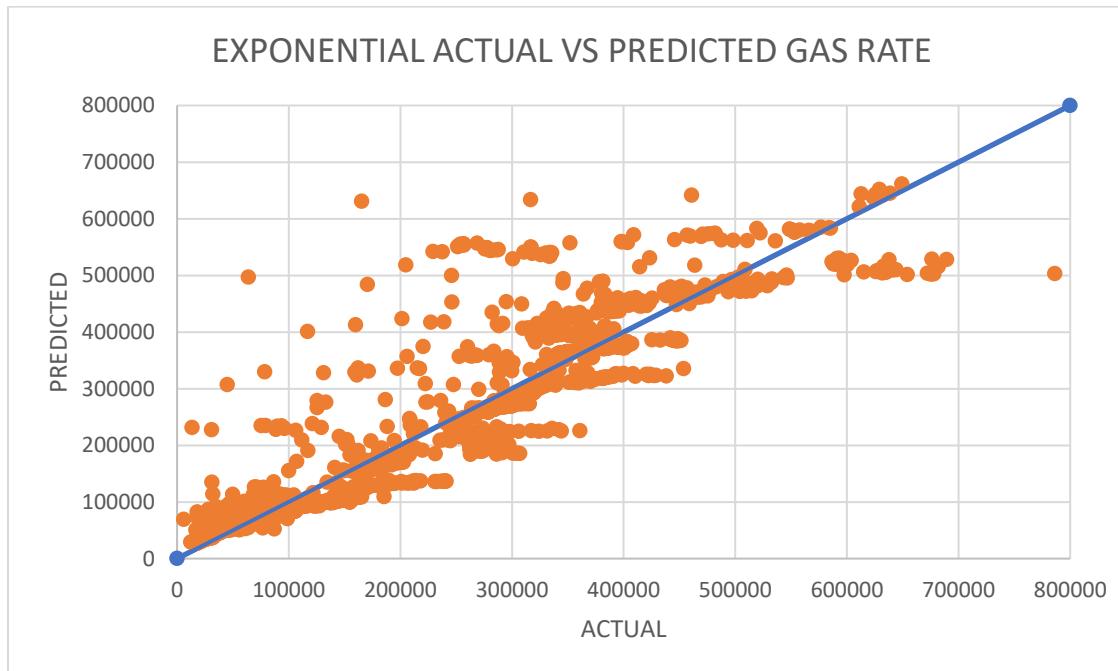
Sai số toàn phương trung bình	RMSE	57071,068
Sai số tuyệt đối trung bình	MAE	33220,225
Hệ số xác định	R ²	0,9299542



Hình PL 12: Đồ thị dự báo lưu lượng khai thác khí theo thời gian trong phương pháp Exponential – DCA



Hình PL 13: Đồ thị dự báo lưu lượng khai thác khí tích lũy theo thời gian trong phương pháp Exponential – DCA



Hình PL 14: Giá trị khai thác khí dự báo so với giá trị khai thác khí thực tế trong phương pháp Exponential - DCA

PHỤ LỤC B

DỰ BÁO KHAI THÁC DẦU KHÍ MỎ VOLVE BẰNG PHƯƠNG PHÁP MẠNG NƠ – RON NHÂN TẠO TRÊN PHẦN MỀM MATLAB

1. Xây dựng mô hình mạng nơ – ron nhân tạo

Bước 1: Xử lý số liệu đầu vào

Để xây dựng mô hình mạng nơ – ron nhân tạo dự báo khai thác mỏ Volve, ta cần chuẩn bị một bộ dữ liệu đầu vào gồm các thông số: thời gian (STEP), sản lượng dầu (OIL), sản lượng khí (GAS) . Các dữ liệu này được chuẩn bị sẵn ở một tập tin Excel.

Trong giao diện của phần mềm MATLAB, tại cửa sổ Workspace, tạo 2 ma trận Input và Target lần lượt là ma trận đầu vào và ma trận mục tiêu để huấn luyện mạng ANN.

Chọn vào ma trận Input, cửa sổ Variables của Input hiện ra, tại đây copy dữ liệu từ tập tin Excel vào. Chuyển tất cả các vectơ cột thành vectơ hàng bằng cách bôi đen toàn bộ vùng dữ liệu bấm chuột phải chọn Transpose Variable. Tương tự nhập dữ liệu mục tiêu Target là giá trị lưu lượng khai thác dầu hoặc khí tại mỗi thời điểm.

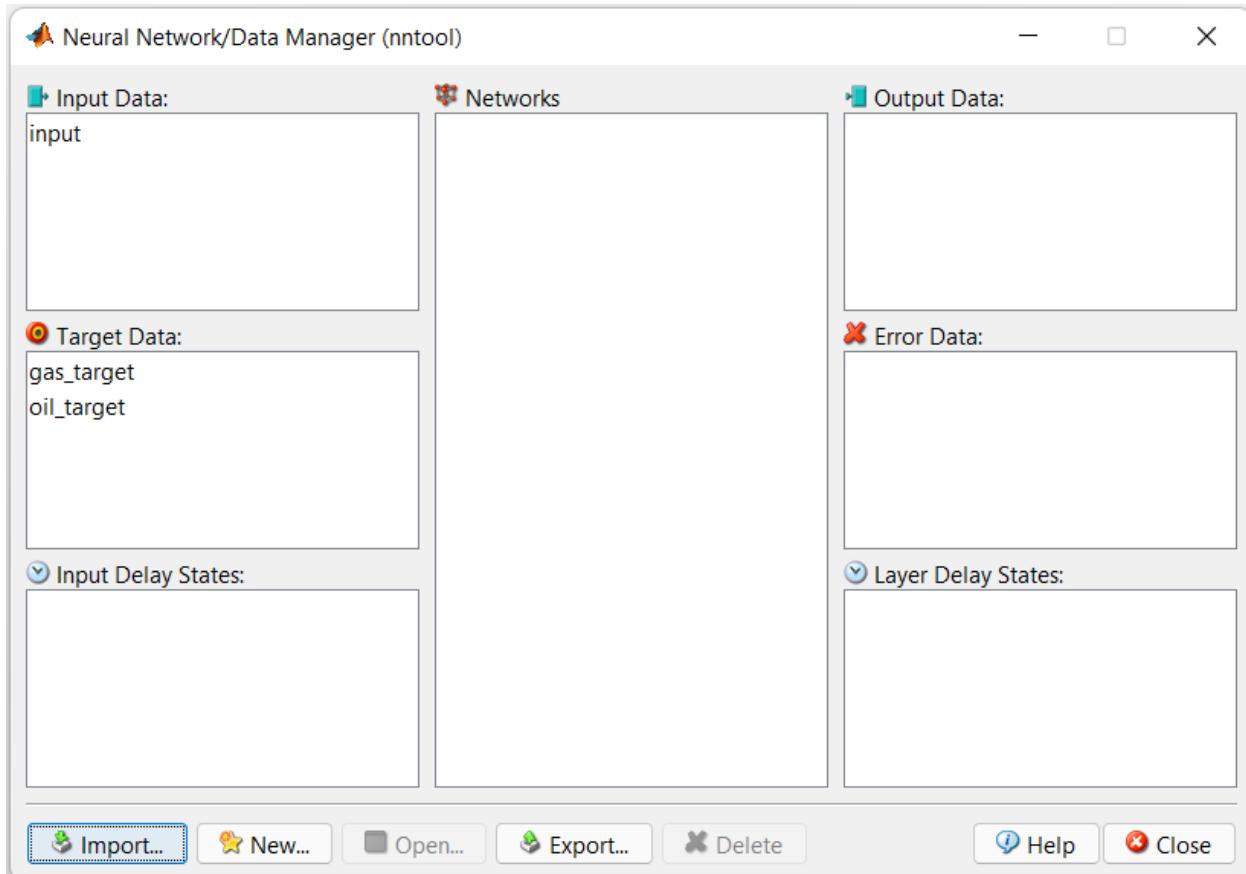
Sau khi chuyển đổi dữ liệu vectơ cột thành vectơ hàng, ma trận Input trở thành 2x2365 (ứng với 2 đặc trưng là STEP và OIL trong trường hợp dự báo sản lượng khí khai thác hoặc GAS trong trường hợp ngược lại và 2356 điểm dữ liệu), ma trận Target trở thành 1x2356 (ứng với biến OIL nếu ta muốn dự báo sản lượng dầu khai thác, tương tự với biến GAS).

Workspace	
Name	Value
gas_target	1x2365 double
input	5x2365 double
oil_target	1x2365 double

Hình PL 15: Các ma trận huấn luyện mạng nơ – ron nhân tạo

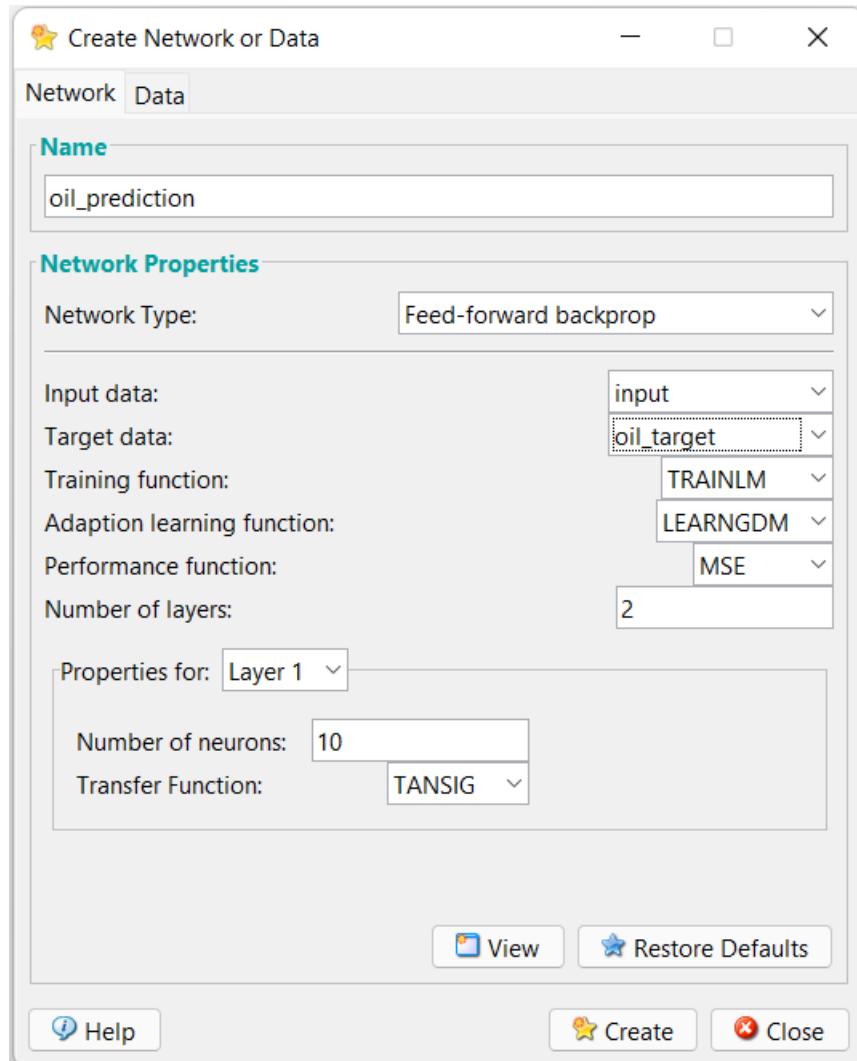
Bước 2: Chọn cấu trúc mạng cho mô hình

Tại Command Window, gõ lệnh “nntool”. Sau đó cửa sổ làm việc Neural Network/Data Manager (nntool) hiện ra.



Hình PL 16: Cửa sổ nntool

Chọn tiếp nút Import và chọn ma trận Input làm Import Data, ma trận Target làm Target Data. Sau đó, chọn vào nút New để mở cửa sổ Create Network or Data.



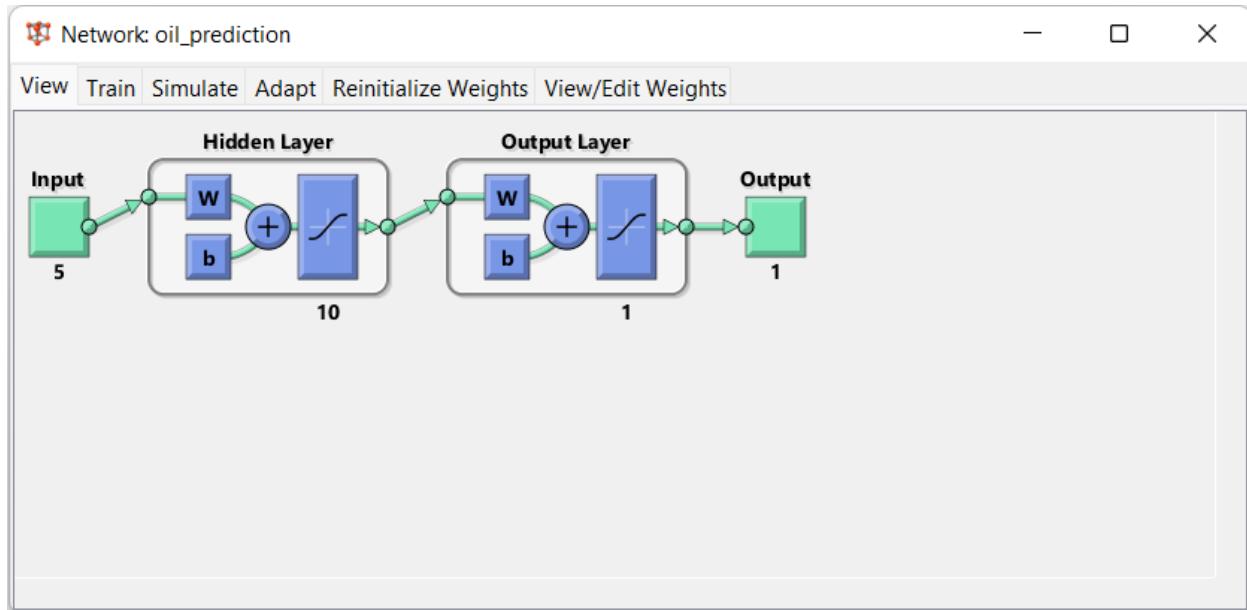
Hình PL 17: Thiết kế cấu trúc mạng ANN

Trong thẻ Network của cửa sổ này, ta có thể đặt tên cho mô hình là Network10 tại ô Name, chọn loại mô hình ANN sử dụng là Feed-forward backprop, chọn Input data và Target data lần lượt là ma trận Input và ma trận Target, Training function chọn TRAINLM ứng với hàm huấn luyện Levenberg-Marquardt, Performance function chọn MSE để trực quan hóa cho kết quả huấn luyện với sai số MSE.

Tiếp theo là vấn đề quyết định chọn số lượng lớp ẩn và số nơ – ron trong mỗi lớp ẩn của mô hình. Trong luận văn này, ta thử lần lượt nhiều mô hình với số nơ – ron và số lớp ẩn tăng dần để tìm được mô hình cho kết quả MSE tốt nhất. Với mỗi Volve, ta chọn số lớp

Ấn là 2, tại ô Number of layers chọn là 2. Chọn Layer 1 là 10 Neuron, hàm kích hoạt cho lớp này là hàm TANSIG, Layer 2 chọn hàm kích hoạt TANSIG.

Như vậy tác giả đã chọn xong cấu trúc cho mạng nơ – ron nhân tạo. Có thể chọn View để xem qua cấu trúc của mô hình như hình PL 6. Sau đó chọn Create để tạo mạng nơron nhân tạo.



Hình PL 18: Cấu trúc mạng nơ – ron thiết kế

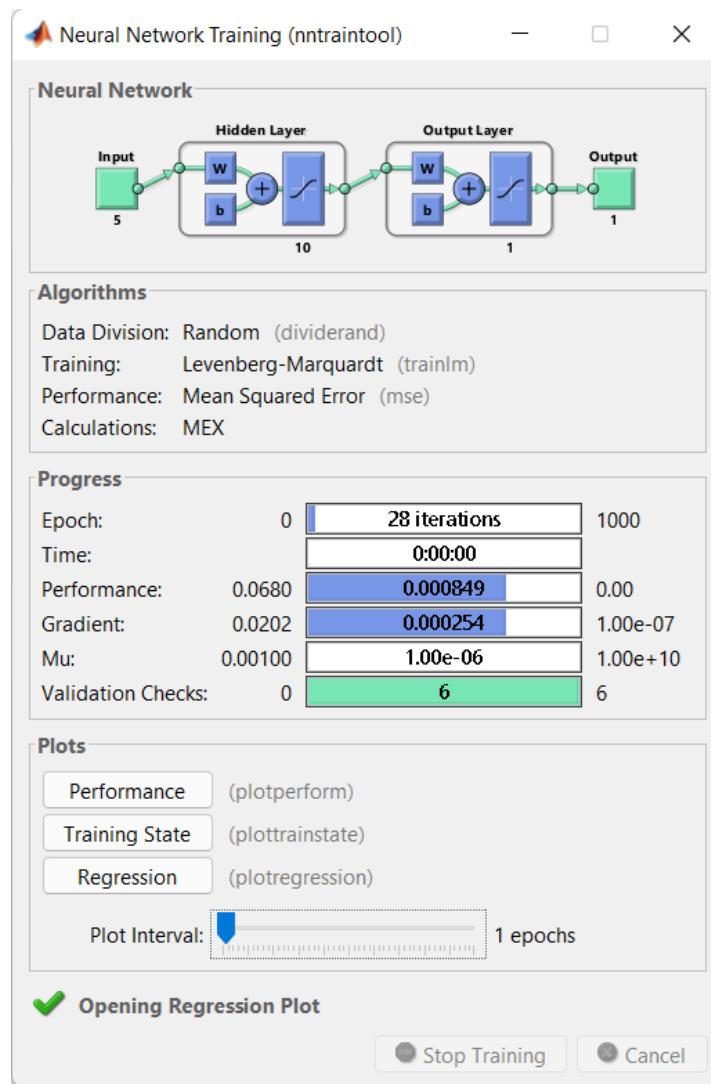
Bước 3: Huấn luyện mô hình

Quay lại cửa sổ làm việc chính của nnTool, chọn Network10 để mở cửa sổ thao tác với mạng nơron. Chọn thẻ Train, nhập Inputs và Targets tương ứng. Cuối cùng chọn Train Network để huấn luyện mô hình.

Bước 4 : Kiểm tra độ chính xác của mô hình

Sau khi hoàn thành bước 3, một cửa sổ Neural Network Training xuất hiện. Cần mất một lúc để quá trình huấn luyện mạng nơ – ron kết thúc. Chọn Performance để vào cửa sổ đồ thị thể hiện sai số của các tập dữ liệu huấn luyện mạng, kiểm tra mạng và kiểm tra chéo.

Quan sát những vòng lặp cuối xem độ ổn định của các sai số để đánh giá độ quá khớp. Nếu mô hình phù hợp có thể đem đi dự báo ứng suất ngang nhỏ nhất.

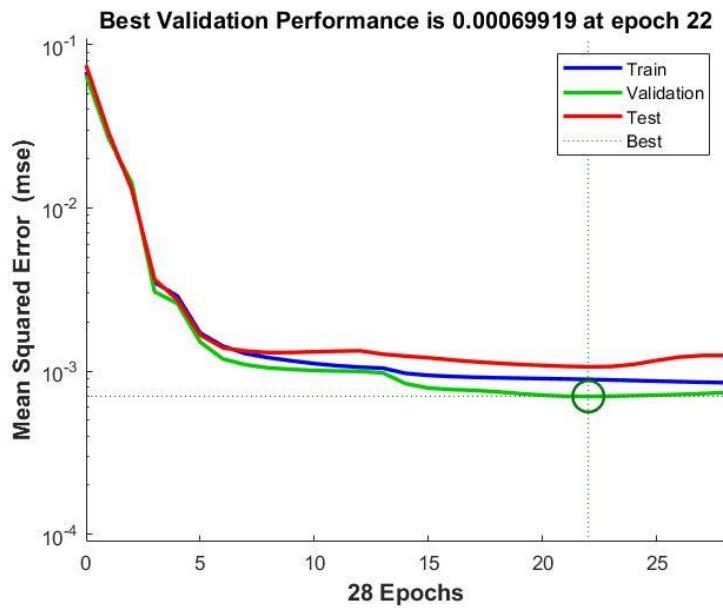


Hình PL 19: Quá trình huấn luyện mạng ANN

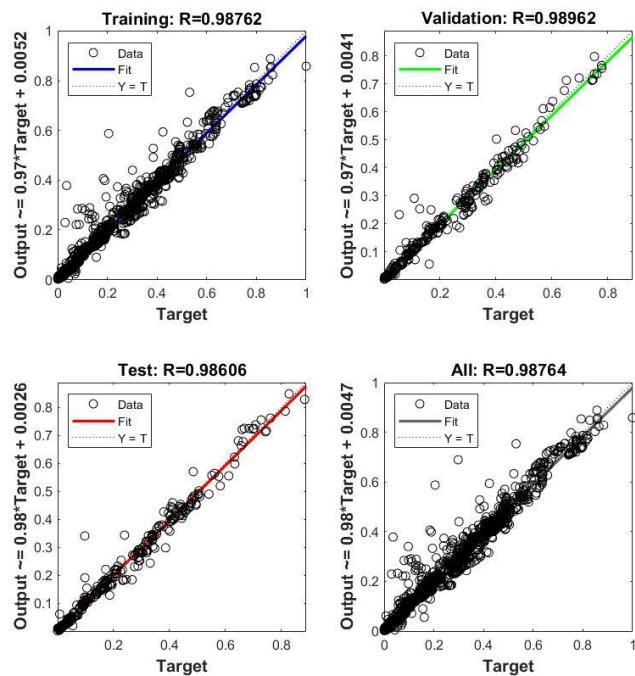
2. Đánh giá kết quả dự báo khai thác dầu khí bằng mạng nơ – ron nhân tạo

Bảng PL 5: Đánh giá mô hình ANN trong dự báo khai thác dầu khí mỏ Volve

Mô hình	RMSE	MAE	R ²	Np	%Np Error
Dự báo sản lượng dầu	165.0261	81.5997	0.9876	2873882	0.001535
Dự báo sản lượng khí	25672.6881	13099.8853	0.9852	423769813	0.006678



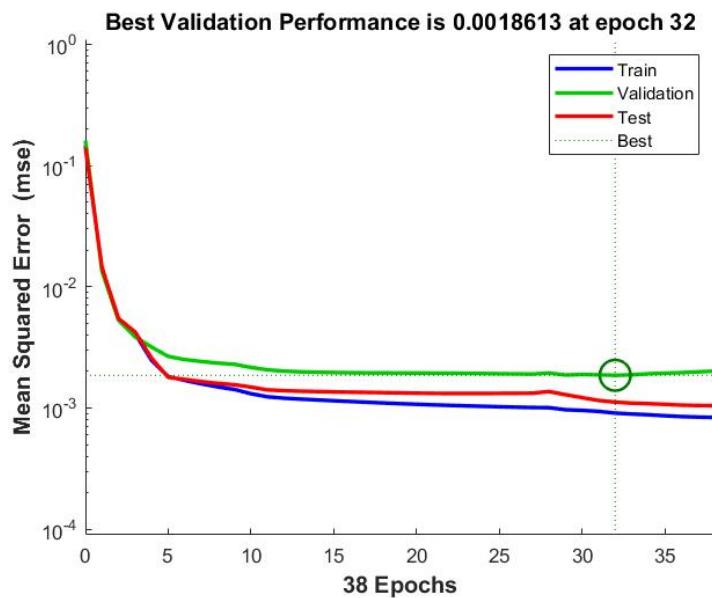
Hình PL 20: Biểu đồ sai số hội tụ của quá trình huấn luyện mạng trong dự báo sản lượng dầu khai thác



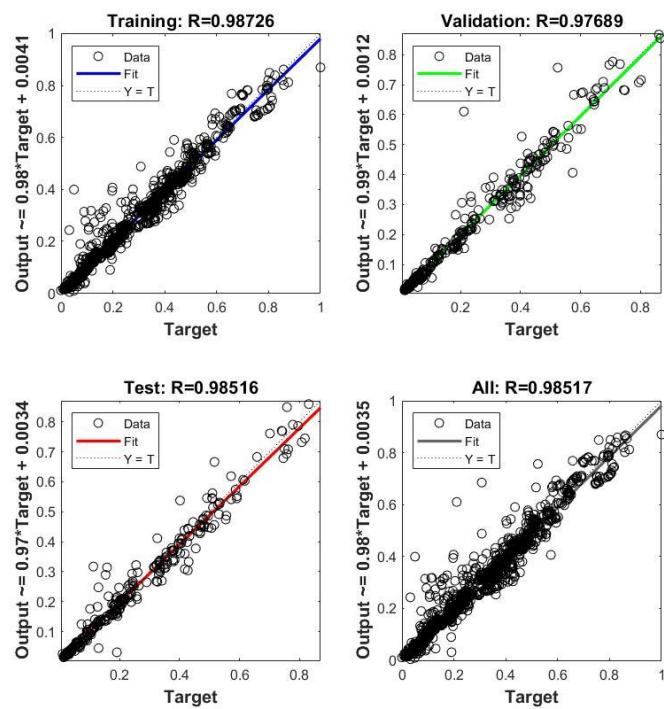
Hình PL 21: Biểu đồ hồi quy của các tập huấn luyện, xác thực và kiểm tra trong dự báo sản lượng dầu khai thác

Bảng PL 6: Kết quả dự báo khai thác dầu mỏ Volve bằng phương pháp ANN

STEP	OIL	predictedOIL	invertOIL	Model Error	Squared Error	Absolute Error
353	4535,4300	0,7876	4466,6416	0,0124	4731,8397	68,7884
365	4379,8800	0,7839	4446,2879	-0,0120	4410,0063	66,4079
374	4509,0700	0,7692	4364,8211	0,0260	20807,7517	144,2489
376	4319,0200	0,7620	4324,6626	-0,0010	31,8385	5,6426
377	4417,6600	0,7718	4379,4110	0,0069	1462,9852	38,2490
379	3226,6100	0,5851	3344,0966	-0,0212	13803,0977	117,4866
380	4411,9000	0,7655	4344,2037	0,0122	4582,7951	67,6963
381	4376,9100	0,7672	4353,5572	0,0042	545,3551	23,3528
382	4417,9100	0,7558	4290,6624	0,0230	16191,9546	127,2476
383	4396,7400	0,7615	4322,2115	0,0134	5554,4928	74,5285
384	4381,0900	0,7599	4313,5122	0,0122	4566,7634	67,5778
389	2208,9600	0,3807	2211,2160	-0,0004	5,0896	2,2560
393	1185,0500	0,2360	1408,5607	-0,0403	49957,0352	223,5107
406	4379,4700	0,7206	4095,5347	0,0512	80619,2680	283,9353
457	4211,2500	0,7089	4030,6063	0,0326	32632,1368	180,6437
459	4237,9400	0,7169	4074,6102	0,0295	26676,6185	163,3298
460	4244,2300	0,7191	4087,2477	0,0283	24643,4580	156,9823
461	3759,1100	0,7208	4096,6441	-0,0609	113929,2763	337,5341
...
3053	103,2100	0,0015	108,8766	-0,0010	32,1107	5,6666
3054	103,2700	0,0015	108,8660	-0,0010	31,3148	5,5960
3055	103,3700	0,0016	108,9415	-0,0010	31,0418	5,5715
3056	102,5700	0,0015	108,9082	-0,0011	40,1722	6,3382
3057	102,8500	0,0015	108,8960	-0,0011	36,5538	6,0460
3058	101,4600	0,0016	108,9862	-0,0014	56,6434	7,5262
3059	102,2600	0,0016	109,0274	-0,0012	45,7982	6,7674
3060	102,3800	0,0016	108,9701	-0,0012	43,4293	6,5901
3061	101,4700	0,0016	109,0001	-0,0014	56,7018	7,5301
3062	102,8900	0,0016	108,9990	-0,0011	37,3200	6,1090
3063	101,5800	0,0016	108,9915	-0,0013	54,9310	7,4115
3064	102,4300	0,0016	109,0109	-0,0012	43,3079	6,5809
3065	100,6700	0,0016	108,9978	-0,0015	69,3522	8,3278
3066	101,8800	0,0016	109,0020	-0,0013	50,7228	7,1220
3068	106,1900	0,0074	141,5818	-0,0064	1252,5811	35,3918
3069	106,3000	0,0094	152,3622	-0,0083	2121,7234	46,0622
3070	102,0900	0,0046	125,9504	-0,0043	569,3170	23,8604
3071	113,3800	0,0094	152,5876	-0,0071	1537,2334	39,2076
3072	108,8400	0,0067	137,2653	-0,0051	807,9996	28,4253
3073	113,8400	0,0095	152,9086	-0,0070	1526,3530	39,0686



Hình PL 22: Biểu đồ sai số hội tụ của quá trình huấn luyện mạng trong dự báo sản lượng khí khai thác



Hình PL 23: Biểu đồ hồi quy của các tập huấn luyện, xác thực và kiểm tra trong dự báo sản lượng khí khai thác

Bảng PL 7: Kết quả dự báo khai thác khí mỏ Volve bằng phương pháp ANN

STEP	GAS	predictedGAS	invertGAS	Model Error	Squared Error	Absolute Error
353	649388,0700	0,8396	661119,3751	-0,0150	137623520,2065	11731,3051
365	629307,3400	0,8215	647012,4149	-0,0227	313469677,5515	17705,0749
374	638750,1700	0,8139	641106,7911	-0,0030	5553663,1915	2356,6211
376	612912,6200	0,7904	622750,8641	-0,0126	96791046,7095	9838,2441
377	625514,0100	0,8005	630616,8578	-0,0065	26039055,5262	5102,8478
379	460948,0100	0,6490	512428,4510	-0,0660	2650235808,3230	51480,4410
380	628668,2700	0,7919	623930,1901	0,0061	22449401,4844	4738,0799
381	625510,2500	0,7904	622779,4505	0,0035	7457266,1512	2730,7995
382	626562,2100	0,7861	619418,5134	0,0092	51032401,5366	7143,6966
383	628354,1200	0,7862	619496,5700	0,0114	78456191,8254	8857,5500
384	623678,1600	0,7817	615955,4053	0,0099	59640940,0124	7722,7547
389	316638,2800	0,3088	246927,7520	0,0893	4859557716,2972	69710,5280
393	165437,3500	0,3062	244880,0845	-0,1018	6311148060,9412	79442,7345
406	611263,1100	0,7842	617916,0531	-0,0085	44261651,5209	6652,9431
457	576830,9500	0,7685	605667,5958	-0,0370	831552140,1792	28836,6458
459	583860,5100	0,7699	606747,5024	-0,0293	523814422,1379	22886,9924
460	585468,5400	0,7687	605821,3251	-0,0261	414235861,7255	20352,7851
461	519316,8800	0,7690	606022,3434	-0,1111	7517837375,1097	86705,4634

3053	16511,0300	0,0158	18252,9599	-0,0022	3034319,7173	1741,9299
3054	16566,1800	0,0158	18264,8341	-0,0022	2885425,8496	1698,6541
3055	16499,8800	0,0160	18386,3642	-0,0024	3558822,4778	1886,4842
3056	16434,6700	0,0159	18370,4099	-0,0025	3747089,1209	1935,7399
3057	16501,9100	0,0160	18400,5169	-0,0024	3604708,2139	1898,6069
3058	16274,3100	0,0163	18642,4103	-0,0030	5607898,8371	2368,1003
3059	16444,3700	0,0164	18741,9123	-0,0029	5278700,4779	2297,5423
3060	16458,5600	0,0164	18705,7284	-0,0029	5049765,8093	2247,1684
3061	16397,1800	0,0165	18776,0764	-0,0030	5659148,0021	2378,8964
3062	16369,8600	0,0165	18802,6165	-0,0031	5918304,4278	2432,7565
3063	16365,9100	0,0165	18838,2293	-0,0032	6112362,6855	2472,3193
3064	16269,7400	0,0166	18884,5026	-0,0034	6836983,3368	2614,7626
3065	16263,0200	0,0166	18887,7279	-0,0034	6889091,6538	2624,7079
3066	16284,4800	0,0167	18927,7221	-0,0034	6986728,5725	2643,2421
3068	17427,7800	0,0176	19681,5035	-0,0029	5079269,6438	2253,7235
3069	17541,2000	0,0192	20898,1865	-0,0043	11269358,1460	3356,9865
3070	16681,2900	0,0202	21681,4306	-0,0064	25001406,0292	5000,1406
3071	18753,1200	0,0192	20902,3168	-0,0028	4619046,9766	2149,1968
3072	17979,2800	0,0195	21170,0025	-0,0041	10180709,9391	3190,7225
3073	18543,7600	0,0196	21262,9838	-0,0035	7394178,1418	2719,2238